The Predictive Utility of Generalized Expected Utility Theories
Author(s): David W. Harless and Colin F. Camerer
Source: *Econometrica*, Vol. 62, No. 6 (Nov., 1994), pp. 1251-1289
Published by: The Econometric Society
Stable URL: http://www.jstor.org/stable/2951749
Accessed: 11/02/2011 13:13

Econometrica, Vol. 62, No. 6 (November, 1994), 1251–1289

THE PREDICTIVE UTILITY OF GENERALIZED EXPECTED UTILITY THEORIES

BY DAVID W. HARLESS AND COLIN F. CAMERER[1]

Many alternative theories have been proposed to explain violations of expected utility (EU) theory observed in experiments. Several recent studies test some of these alternative theories against each other. Formal tests used to judge the theories usually count the number of responses consistent with the theory, ignoring systematic variation in responses that are inconsistent. We develop a maximum-likelihood estimation method which uses all the information in the data, creates test statistics that can be aggregated across studies, and enables one to judge the predictive utility—the fit and parsimony—of utility theories. Analyses of 23 data sets, using several thousand choices, suggest a menu of theories which sacrifice the least parsimony for the biggest improvement in fit. The menu is: mixed fanning, prospect theory, EU, and expected value. Which theories are best is highly sensitive to whether gambles in a pair have the same support (EU fits better) or not (EU fits poorly). Our method may have application to other domains in which various theories predict different subsets of choices (e.g., refinements of Nash equilibrium in noncooperative games).

KEYWORDS: Expected utility theory, non-expected utility theory, prospect theory, model selection, Allais paradox.

DISSATISFACTION WITH THE EMPIRICAL ACCURACY of expected utility (EU) theory has led many theorists to develop generalizations of EU. The development of alternatives to EU, in turn, has led to a vigorous new round of experiments testing the empirical validity of the new theories against each other and against EU (Battalio, Kagel, and Jiranyakul (1990), Camerer (1989, 1992), Chew and Waller (1986), Conlisk (1989), Harless (1992), Prelec (1990), Sopher and Gigliotti (1990), Starmer and Sugden (1989)). The experiments test robustness of previously observed EU violations (Allais (1953), Kahneman and Tversky (1979)) and test the accuracy of predictions in new domains.

The recent studies are informative and useful—for example, recent results have already guided development of some new theories[2]—but there is still substantial confusion about what the new studies say. For example, the Chew and Waller (1986) data have been cited as supporting weighted EU theory (by Chew and Waller), as supporting the "fanning out" hypothesis (by Machina (1987)), and as supporting a mixture of fanning out and "fanning in" (by Conlisk (1989)).

In this paper we show that confusion about the results of the new studies can be largely resolved by more powerful statistical tests. Our paper makes three contributions: We present new tests, which gain power by using all the information available in patterns of observed choices (most earlier tests threw away

[1] Thanks to John Conlisk, Dave Grether, Bill Neilson, Nat Wilcox, and a co-editor and two anonymous referees for helpful comments, to Drazen Prelec and John Kagel for supplying their data, and especially to Teck-Hua Ho for collaboration in the project's early stages. Camerer's work was supported by NSF Grant SES-90-23531 and by the Russell Sage Foundation, where he visited during the 1991–1992 academic year.

[2] See Neilson (1992a, 1992b), Chew, Epstein, and Segal (1991).

some important information); the test statistics we derive can be added across studies, enabling us to aggregate data (nearly 8,000 choices) and increasing power further; and we give a method for trading off fit and parsimony of various theories. Hence, our work explores the predictive utility—fit and parsimony—of various utility theories.

The result is a menu of theories. Researchers can pick a theory from the menu, depending on the price they are willing to pay (in poorer fit) for added parsimony. Aggregating across all studies, the menu is: mixed fanning, prospect theory, EU, and expected value (EV); but the results are sensitive to the domain of gambles.

The paper proceeds as follows. The next section illustrates our method, and predictions of several generalized EU theories, with one study. Section 2 reviews the results from several choice studies. Section 3 aggregates the results from 23 data sets and 2,000 choice patterns, and describes a method for trading off fit and parsimony. In Section 4 we draw conclusions.

1. ILLUSTRATION OF OUR MAXIMUM-LIKELIHOOD ANALYSIS

The study by Battalio, Kagel and Jiranyakul (1990), one of several we include in our analyses, will illustrate our method and the predictions of several generalized utility theories. In one part of their study, subjects chose one lottery (or expressed indifference) out of each of three pairs. Each pair consisted of one lottery, denoted S for "safer," and a mean-preserving spread of S, denoted R for "riskier." The pairs were:

Pair 1: $S1 = (-\$20,.6;-\$12,.4)$	$R1 = (-\$20,.84;\$0,.16)$
Pair 2: $S2 = (-\$12)$	$R2 = (-\$20,.6;\$0,.4)$
Pair 3: $S3 = (-\$12,.2;\$0,.8)$	$R3 = (-\$20,.12;\$0,.88)$

Figure 1 shows the three pairs in a unit triangle diagram (Marschak (1950), Machina (1982)). In the diagram, each lottery is plotted as a point along the

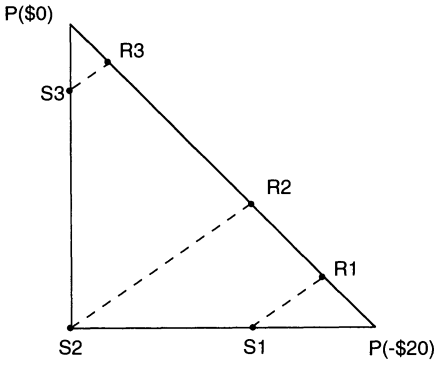


FIGURE 1.—Unit triangle example.

TABLE I

CONSISTENT PATTERNS FOR BATTALIO, KAGEL, AND JIRANYAKUL REAL LOSSES

Pattern · 123	Observed Frequency	EU	Fan Out	Fan In	MF	RD-cave gIE	RD-cave fDE	RD-vex gDE	RD-vex fIE	PT
SSS	7	X	X	X	X	X	X	X	X	
SSR	1			X	X	X	X	X		
SRS	1							X	X	
SRR	1			X	X		X	X	X	
RSS	3		X		X	X	X		X	
RSR	0				X	X	X			
RRS	8		X		X	X		X	X	X
RRR	7	X	X	X	X	X	X	X	X	X

$p(-\$20)$ and $p(0)$ axes. Notice that the pairs are related in a particular geometric way: The lines connecting the lotteries in each pair are parallel. Furthermore, two of the pairs have a common ratio of outcome probabilities—for example, $p(-\$12)/p(-\$20)$ equals $1/.6 = 5/3$ in pair 2, and $.2/.12 = 5/3$ in pair 3. The two pairs form a "common ratio" problem. (In pair 1, the ratio of the differences in outcome probabilities between lotteries, $(.4 - 0)/(.84 - .6)$, has the ratio 5/3 too.)

Most choice theories do not predict precisely whether people will pick S or R in each pair. Instead, theories restrict patterns of choices across pairs. For example, a person who obeys EU judges gambles by the expectation of the utilities of their outcomes. Thus, a person who obeys EU and prefers $S2$ to $R2$ (denoted $S2 \succ R2$) reveals that $u(-\$12) > .6u(-\$20)$ (setting $u(\$0) = 0$ for simplicity). But $u(-\$12) > .6u(-\$20)$ implies $.2u(-\$12) > .12u(-\$20)$, which predicts $S3 \succ R3$. By a similar calculation, $S1 \succ R1$. EU therefore predicts that people who choose S in one pair will choose S in the other pairs too, so EU allows the pattern denoted *SSS*—the choice of $S1$, $S2$, and $S3$. Alternatively, an EU-maximizer with $R2 \succ S2$ must prefer $R1$ and $R3$ to $S1$ and $S3$, so EU allows the pattern *RRR*, but not the other six possible patterns.

The patterns allowed by each theory in the BKJ study are shown in Table I (marked by X's). We review the predictions of each theory briefly. Much more detail is available in Machina (1982, 1987), Fishburn (1988), and Camerer (1989, 1992).

EU: As described above, EU predicts patterns *SSS* or *RRR* in Table I. Graphically, EU requires the indifference curves that connect sets of equally-preferred gambles to be parallel straight lines. That implies preference for the S lottery in each pair, or the R lottery in each pair.

Fanning out: Machina (1982) proposed a generalization of EU in which Frechet differentiability of a preference functional guaranteed that similar lotteries could be approximately ranked by the EU of a "local" utility function. (In EU, the local utility functions are all the same.) He also suggested that many violations of EU could be explained by the hypothesis that as the lotteries being ranked become better (in the sense of stochastically-dominating improvements),

local utility functions become more concave (reflecting greater local risk-aversion). Graphically, Machina's hypothesis implies that indifference curves "fan out:" curves become steeper as one moves in the direction of increasing preference, from the lower right hand corner to the upper left hand corner. Besides the EU-conforming patterns *SSS* and *RRR*, fanning out allows any patterns in which preferences switch from R to S from pairs 1 to 3, viz., patterns *RSS* and *RRS*.

Fanning in: The opposite of fanning out is "fanning in," the tendency of indifference curves to become flatter, not steeper, in the direction of increasing preference. There is little *a priori* evidence suggesting fanning in, but we consider it for completeness. Fanning in allows patterns *SSR* and *SRR* (along with the EU patterns).

Mixed fan (MF): There is some evidence that indifference curves fan out for less favorable lotteries (like pair 1) and fan in for more favorable ones (like pair 3), suggesting a hybrid "mixed fan" hypothesis (cf. Neilson (1992a)) in which the direction of fanning switches within a triangle diagram. The point at which fanning switches from out to in (moving to the northwest) might lie outside the space of choices, so both fanning out patterns and fanning in patterns are consistent with MF. Mixed fanning also allows a pattern that neither fanning out nor fanning in allows, viz., fanning out between pairs 1 and 2, and fanning in between pairs 2 and 3, the pattern *RSR*. The only pattern which is excluded is *SRS*.

EU with rank-dependent weights (RD): There are several generalizations of EU in which outcome utilities are weighted "rank-dependently" or "cumulatively" (Quiggin (1982), Yaari (1987), Chew, Karni, and Safra (1987), Green and Jullien (1988), Segal (1987, 1989), Tversky and Kahneman (1992)). In most of these theories, a decumulative distribution function (one minus the cumulative distribution function) is transformed by a continuous, monotonic function $g(p)$, with $g(0) = 0$ and $g(1) = 1$; outcomes are weighted by differences or differentials of the transformed decumulative. If $g(p)$ is *convex* then high-ranked outcomes are underweighted and unit triangle indifference curves are *concave* (denoted RD-cave); curves also fan out along the base of the triangle, and fan in along the left side. If $g(p)$ is *concave* then high-ranked outcomes are overweighted and indifference curves are *convex* (denoted RD-vex); curves fan in along the base of the triangle, and fan out along the left side. If $g(p) = p$ then each outcome is simply weighted by its probability, as in EU.

RD theories do not make precise predictions about choices in the BKJ study unless further restrictions are placed on the shape of $g(p)$. Segal (1987) showed that if $g(p)$ is convex (indifference curves are concave) and has increasing elasticity (i.e., $pg'(p)/g(p)$ is increasing in p) then indifference curves will exhibit a common ratio effect: fanning out in the southeast portion of the triangle (*SSx*, *RRx*, *RSx* allowed; *SRx* not allowed). We denote predictions when indifference curves are concave (and $g(p)$ is convex with *i*ncreasing *e*lasticity) as RD-cave *g*IE. The theory makes different predictions when indifference curves are convex and $g(p)$ is concave (denoted RD-vex) and when

cumulative probabilities are weighted instead of decumulatives (denoted by labeling the weighting function f rather than g). The predictions of four variants of RD with elasticity conditions are shown in Table I. Quiggin's (1982) original form of rank-dependent expected utility, called "anticipated utility," presumed $f(.5) = .5$ with $f(p)$ concave below $.5(f(p) > p)$ and convex above $.5(f(p) < p)$ (see also Quiggin (1993)). This form excludes no patterns in 8 of the 23 studies we consider later, so we say nothing more about it except in footnote 23.

Prospect theory (PT): Kahneman and Tversky (1979) proposed a descriptive theory embodying several empirical departures from EU. We test an extremely simplified form of prospect theory which incorporates several of its key features: The value function, or utility function over riskless amounts, is assumed to have a reference point of \$0 (i.e., $v(\$0) = 0$) and to be strictly concave for gains and strictly convex for losses (exhibiting a "reflection effect"); probabilities are assumed to be transformed by a decision weight function $\pi(p)$;[3] and lotteries are ranked by the sum of their weighted outcome values.[4]

All the predictions we derive based on prospect theory require only reflection of the value function and certain general properties of decision weights as hypothesized in Kahneman and Tversky (1979). The properties of the probability transformation function $\pi(p)$ we use in making predictions are subcertainty ($\pi(p) + \pi(1-p) < 1$), subproportionality ($\pi(rp)/\pi(rq) > \pi(p)/\pi(q)$, for $p < q$ and $0 < r < 1$), and convexity of π for probabilities above .01, which allows for overweighting small probabilities and underweighting larger probabilities (but we assume only that the crossover point occurs somewhere between .1 and .3). In the BKJ study prospect theory predicts $v(S2) = v(-12)$ and $v(R2) = \pi(.6)v(-20)$. Convexity of $v(x)$ for losses implies $v(-12) < .6v(-20)$; underweighting of high probabilities implies $\pi(.6) < .6$. Together they imply $v(-12) < \pi(.6)v(-20)$, predicting a preference for $R2$ over $S2$. Prospect theory also predicts a preference for $R1$ over $S1$, but makes no prediction about choices in pair 3.[5] Here, the theory as we have characterized it allows only the two patterns *RRS* and *RRR*, so it is just as parsimonious as EU.

Additional theories: There are many other choice theories besides those whose predictions are shown in Table I. We consider some and neglect others. The historical predecessor to EU, expected value maximization (EV), predicted that people would choose lotteries according to expected value. In some studies,

[3] Tversky and Kahneman (1992) show how to extend prospect theory in several ways, including cumulative weighting as in the rank-dependent theories.

[4] In addition, "irregular lotteries," which have only positive or only negative outcomes, are valued by segregating the certain outcome from the uncertain part.

[5] $S1$ chosen over $R1$ implies $(1 - \pi(.6))v(-12) + \pi(.6)v(-20) > \pi(.84)v(-20)$; the convexity of the value function implies $1 - \pi(.6) > (1/.6)(\pi(.84) - \pi(.6))$ which contradicts the assumption that π is convex. Prospect theory makes no prediction about choices in pair 3 because $S3 \gtrless R3$ as $\pi(.12) \gtrless .6\pi(.2)$. Convexity of the value function for losses means PT predicts preference for the riskier lottery when lotteries are mean-preserving risk spreads except when the decision weight function overweights small probabilities (such as the .12 probability of $-\$20$ in $R3$).

including the BKJ study illustrated by Table I, lotteries in a pair had the same expected value. Then we took EV to be identical to EU.[6]

If the local utility function in generalized utility is constant along an indifference curve, then "implicit EU" (IEU) results (Dekel (1986), Chew (1989)). IEU predicts linear indifference curves, thus satisfying the betweenness axiom, a weakened form of independence. (Betweenness requires that a reduced-form probability mixture of any two lotteries should not be worse or better than both; the mixture should lie between them in preference.) IEU is only tested in studies which ask subjects to choose between lotteries in two or more pairs which lie on the same line in the triangle diagram. In the BKJ study IEU allows any pattern (since no two pairs lie on the same line).

In "weighted utility" theory (WEU), a special case of IEU, lottery utilities are computed by multiplying an outcome's utility by its probability and by a normalized weighting function which depends on the outcome (Chew and MacCrimmon (1979), Chew (1983), Fishburn (1982, 1983)). In all the studies we consider, WEU is the same as combining either fanning out and fanning in (depending on the shape of the weighting function) with linearity of indifference curves; we denote these brands of WEU as WEU-out and WEU-in. In some studies, like BKJ, the predictions of WEU-out (WEU-in) are the same as those of fanning out (in).

Combining mixed fanning with linear indifference curves yields a hybrid we call "linear mixed fan" (LMF). This theory was suggested by Neilson (1992a). Gul's (1991) disappointment-based theory is slightly more restrictive but observationally equivalent to LMF in all the studies we review. (However, a special study could be designed to separate LMF and Gul's theory.)

Theories *not* included in the tables below include lottery-dependent utility (Becker and Sarin (1987)), ordinal utility (Green and Jullien (1988)), prospective reference theory[7] (Viscusi (1989)), combinations of rank-dependent and weighted utility (Chew and Epstein (1990)), and the cumulative extension of prospect theory proposed recently (Tversky and Kahneman (1992)). We address some of these theories in the footnotes and Section 3. Others may be easily tested using our method after the hard work of determining which choice patterns the theories allow is finished.

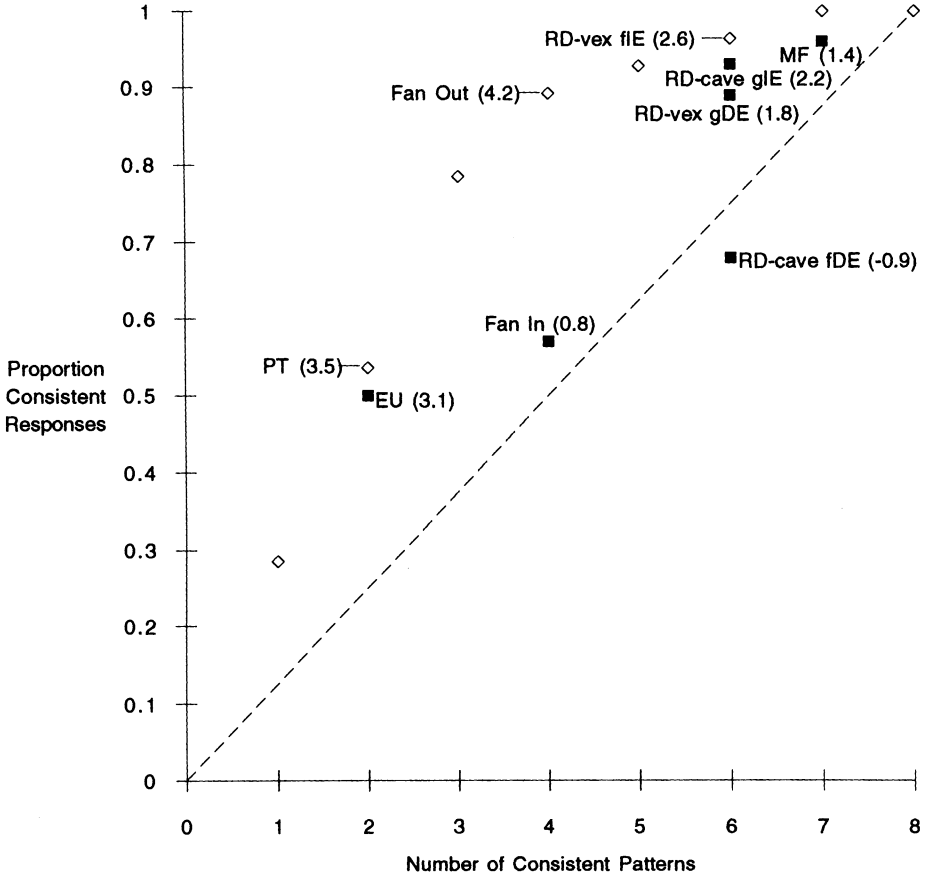
[6] Alternatively, when one gamble is a mean-preserving spread of another, one could interpret EV to imply that subjects are indifferent between the gambles in each pair. Under that interpretation, each pattern is equally likely (the EV maximizer chooses by flipping a fair coin). Tests of that restriction are reported in footnote 27. Another approach allows each indifferent pair to have a different choice proportion, then estimate the likelihood-maximizing proportions from the data. (The coin-flip approach restricts the proportions to be .5.) But this approach allows EV to fit too well: if the choice-proportion parameters are allowed to differ across pairs, EV, so interpreted, may "explain" fanning out or fanning in. We could allow other theories (which have EV as a special case) the same luxury of extra choice-proportion parameters, but then we quickly run out of degrees of freedom in many data sets.

[7] We have not included prospective reference theory in the main analysis because we only study experiments with three, four, and five pairwise choices (see footnote 21), and none of these experiments adequately tests the predictions of prospective reference theory. For tests of the specific predictions of prospective reference theory with two pairwise choices, see Harless (1993).

Graphical Comparison of Theories

A simple, appealing way to judge a theory is to calculate the fraction of observed patterns that are consistent with the theory. EU allows the patterns *SSS* and *RRR*. Table I shows that seven people picked *SSS* and seven picked *RRR*, so 14 of 28 (50%) chose as predicted by EU. The four patterns allowed by fanning out were picked by 25 of 28 subjects (89%).

An obvious drawback of this method is that theories which allow more patterns will always have a higher proportion of consistent choices. A simple test which puts theories on more equal footing compares the proportion of consistent choices with the proportion of patterns allowed (i.e., the proportion of choices that would be consistent if choices were actually made randomly). For example, EU patterns capture 50% of the choices, but allow just 2 of 8 possible patterns (25%). A z test measures the likelihood of such accurate prediction if



Z statistics (in parentheses) test each theory against the random choice null hypothesis.

FIGURE 2.—Battalio, Kagel, and Jinanyakul: Real losses, Series 1.

choices were made randomly (e.g., Chew and Waller (1986)). For EU, the z statistic is $(.50 - .25)/[(.25)(.75)/28)]^{1/2}$, or $z = 3.1$ ($p = .001$). For fanning out, which allows four of eight possible patterns (50%) and explains 89% of the choices, $z = 4.2(p = 1.3E - 05)$.

There are many plausible ways to compare the accuracy of theories (e.g., comparing their z statistics). We developed a graphical method of displaying accuracy which permits easy ranking of theories by several criteria. Figure 2 gives an example using the BKJ data.

Figure 2 shows the proportion of consistent responses (y axis) and the number of patterns allowed (x axis), for several theories. The open diamonds represent the "data frontier:" the highest possible consistent proportions for each number of consistent patterns. For example, the three most common patterns, *RRS*, *RRR*, and *SSS* were chosen by 29%, 25%, and 25% of the subjects, respectively. The best one-pattern theory would have only 29% consistent. The best two-pattern theory would have 54% consistent, and so on. (Note that the data frontier is always (weakly) concave.) The data frontier therefore shows the best a theory can possibly do. Points representing different theories must always lie below the data frontier or lie right on it (as prospect theory, fanning out, and RD-vex-*f* IE do, in Figure 2). The hatched "random choice line" represents the proportions consistent that would result if people chose randomly.

A look at Figure 2 suggests some visual ways to judge theories. Good theories should be close to the data frontier, and far from the random choice line. The z statistic (shown in parentheses in Figure 2) gives a formal measure of how far each point is from the random choice line; fanning out does best by that criterion ($z = 4.2$). The difference between the proportion consistent and the proportion of patterns allowed, a measure advocated by Selten (1991), is the vertical (or horizontal) distance from the random choice line; fanning out does best by that criterion also.[8] The ratio measure, the proportion of consistent choices *per pattern*, is measured by the slope of the line connecting each theory point to the origin; prospect theory is best by that measure (27% per pattern). An opposite measure is the proportion of *in*consistent choices per *in*consistent pattern (sometimes called the "outside ratio"), measured by the slope of the line connecting each theory point to the upper right corner. A good theory makes the outside ratio low; RD-vex-*f* IE is best by that measure.

The various criteria reward theories for different kinds of predictive accuracy. The ratio statistic (slope from origin) rewards more parsimonious theories which capture the most common pattern(s). The outside ratio (slope from upper right corner) rewards broader theories which exclude uncommon patterns.

[8] The z statistic and the difference measure are closely related because the z statistic is simply the difference measure divided by $(p(1-p)/n)^{1/2}$. Since the number of observed pattern choices n is the same for all the theories, compared to the difference measure the z statistic favors theories with low and high values of p (i.e., theories that predict very few or very many patterns).

Figure 2 shows that many of the generalizations of EU are surprisingly parsimonious and accurate. For example, in this study prospect theory predicts the same number of patterns as EU (two) but it explains more choices and beats EU by all the measures given above. Fanning out permits twice as many patterns as EU, but it accounts for nearly twice as many choices (and beats EU by the measures except ratio). The graph is useful for screening out dominated theories—those which allow the same number of patterns (or more) but have fewer consistent responses than other theories. Dominated theories lie to the lower right of theories which dominate them. In Figure 2, prospect theory dominates EU, fanning out dominates fanning in, RD-vex-gDE and RD-cave-fDE, and RD-vex-fIE dominates mixed fan and the other rank-dependent theories.

Note that we use the terms parsimonious in a very specific sense, to denote the number of patterns a theory allows. However, the number of patterns a theory allows does not necessarily correspond to the number of free parameters or free functions it uses. Theories which appear unparsimonious because they have many additional free parameters or free functions might, with minimal restrictions on those functions, predict relatively few patterns (prospect theory is an example). Contrarily, a theory which has only one free parameter more than EU may allow a wide range of patterns and hence be unparsimonious by our standard (Gul's (1991) one-parameter disappointment-based theory is an example).

Comparing Theories with Maximum-Likelihood Error Rate Analysis

The analyses expressed visually in Figure 2 have two severe shortcomings: First, there is no single compelling measure by which to compare theories. Second, all the criteria throw away information by collapsing the entire distribution of responses into a single number—the proportion of choices consistent with a theory.

Our test overcomes these problems. We characterize a theory as a restriction on the proportions of subjects that have true preferences corresponding to each of the eight patterns. For example, EU permits two types of subjects, a proportion p_1 of consistent risk-averters who prefer *SSS*, and a proportion $1 - p_1$ of consistent risk-preferrers who prefer *RRR*. In previous work (including the studies we reanalyze in this paper, some of which are our own), if a subject were to choose, say, *RRS* or *SRR*, the response was simply counted as inconsistent with EU. The premise of our test is that systematic variation in unpredicted patterns should count against a theory: if many people choose *RRS* and few choose *SRR*, a theory which predicts nobody will choose either should be penalized more heavily.

Penalizing theories for systematic variation in unpredicted patterns requires some allowance for error; otherwise, a single observation of an unpredicted pattern would immediately invalidate a theory. Therefore, we allow the possibil-

ity of erroneous deviations from underlying preferences so we can judge the *degree* of inconsistency of an observation.[9] For example, suppose EU is true—people prefer either *RRR* or *SSS*—but subjects make random errors which are independent and equally likely across the three choices. For those subjects with true preference pattern *RRR*, the patterns which occur because of one error (*SRR*, *RSR*, and *RRS*) should be equally likely, and should be more likely than the two-error patterns (*SSR*, *SRS*, and *RSS*). For those subjects with true preference pattern *SSS*, the patterns which occur because of one error (*RSS*, *SRS*, and *SSR*) should be equally likely and should be more likely than the two error patterns (*SRR*, *RSR*, and *RRS*). Thus, EU can be characterized as a restriction on allowed patterns (*SSS* and *RRR* patterns only), which implies—when error is assumed—that some inconsistent patterns are more likely than others. By assuming a range of true underlying preferences (restricted by the theory) and an error rate, each theory makes interconnected predictions about the relative frequency of *each* consistent and inconsistent pattern. We can then use the entire distribution of choices to judge a theory, rather than simply counting totals of consistent or inconsistent choices, or restricting attention to two choices as previous studies have.[10]

The BKJ data illustrate how our method works. Fanning out allows four types of subjects: Those who choose *SSS*, *RSS*, *RRS*, and *RRR*. Call the proportions of people with each preference $p(SSS)$, $p(RSS)$, $p(RRS)$, and $p(RRR)$. The theory predicts that there are no subjects with true preference for *SSR*, *SRS*, *SRR*, and *RSR*, but those patterns can result if people make errors in expressing true preferences. Errors occur with probability ε, and are independent for each choice. For fanning out, Table II shows the patterns which can result for each of the true preferences for various numbers of errors, and the resulting likelihood function. For example, a *RSS*-type who makes exactly two errors—which happens with probability $p(RSS)\varepsilon^2(1-\varepsilon)$—could choose, *SRS*, *SSR*, or *RRR*. The total probability of choice *RRR* is $p(SSS)\varepsilon^3 + p(RSS)\varepsilon^2(1-\varepsilon) + p(RRS)\varepsilon(1-\varepsilon)^2 + p(RRR)(1-\varepsilon)^3$. For each theory we find values of the true pattern proportions and the error rate (restricted to lie between 0 and 0.5) which maximize the likelihood of the distribution of responses under each theory's restrictions on consistent patterns.

We assume a single error rate for all three choices for several reasons. First, it is a parsimonious and conservative approach to explaining the distribution of choice responses. Many researchers have implicitly adopted independent and equal errors in statistical tests of choice theories with two pairwise choices. We take that underlying model of errors and apply it to data sets with three or more

[9] When indifference curves are convex (i.e., preferences are quasi-concave), what we call "errors" might be expressions of strict preference for randomization (Machina (1985), Crawford (1988)). We show in an unpublished Appendix (available on request) that our tests are equivalent to the proper test when indifference curves are convex.

[10] Conlisk (1989) used a similar error rate with two pairs, but didn't make use of the error rate in his statistical test. Starmer and Sugden (1989) and Lichtenstein and Slovic (1971) incorporated error rates too.

TABLE II

EXAMPLE OF OCCUPATION OF PATTERNS WHEN SUBJECTS MAKE ERRORS: FANNING OUT IN BATTALIO, KAGEL, AND JIRANYAKUL

Consistent Pattern	Zero Errors	One Error	Two Errors	Three Errors
SSS	*SSS*	*RSS, SRS, SSR*	*RRS, RSR, SRR*	*RRR*
RSS	*RSS*	*SSS, RRS, RSR*	*SRS, SSR, RRR*	*SRR*
RRS	*RRS*	*SRS, RSS, RRR*	*SSS, SRR, RSR*	*SSR*
RRR	*RRR*	*SRR, RSR, SRR*	*SSR, SRS, RSS*	*SSS*

Fanning out likelihood function

$$
\begin{aligned}
&[p(SSS)(1-\varepsilon)^3 + p(RSS)\varepsilon(1-\varepsilon)^2 + p(RRS)\varepsilon^2(1-\varepsilon) + p(RRR)\varepsilon^3]^{(\text{frequency } SSS)} \times \\
&[p(SSS)\varepsilon(1-\varepsilon)^2 + p(RSS)\varepsilon^2(1-\varepsilon) + p(RRS)\varepsilon^3 + p(RRR)\varepsilon^2(1-\varepsilon)]^{(\text{frequency } SSR)} \times \\
&[p(SSS)\varepsilon(1-\varepsilon)^2 + p(RSS)\varepsilon^2(1-\varepsilon) + p(RRS)\varepsilon(1-\varepsilon)^2 + p(RRR)\varepsilon^2(1-\varepsilon)]^{(\text{frequency } SRS)} \times \\
&[p(SSS)\varepsilon^2(1-\varepsilon) + p(RSS)\varepsilon^3 + p(RRS)\varepsilon^2(1-\varepsilon) + p(RRR)\varepsilon(1-\varepsilon)^2]^{(\text{frequency } SRR)} \times \\
&[p(SSS)e(1-\varepsilon)^2 + p(RSS)(1-\varepsilon)^3 + p(RRS)\varepsilon(1-\varepsilon)^2 + p(RRR)\varepsilon^2(1-\varepsilon)]^{(\text{frequency } RSS)} \times \\
&[p(SSS)e^2(1-\varepsilon) + p(RSS)\varepsilon(1-\varepsilon)^2 + p(RRS)\varepsilon^2(1-\varepsilon) + p(RRR)\varepsilon(1-\varepsilon)^2]^{(\text{frequency } RSR)} \times \\
&[p(SSS)e^2(1-\varepsilon) + p(RSS)\varepsilon(1-\varepsilon)^2 + p(RRS)(1-\varepsilon)^3 + p(RRR)\varepsilon(1-\varepsilon)^2]^{(\text{frequency } RRS)} \times \\
&[p(SSS)\varepsilon^3 + p(RSS)\varepsilon^2(1-\varepsilon) + p(RRS)\varepsilon(1-\varepsilon)^2 + p(RRR)(1-\varepsilon)^3]^{(\text{frequency } RRR)}.
\end{aligned}
$$

pairwise choices where the same model of errors generates more powerful tests of the choice theories.

Second, allowing error rates to be choice-dependent can lead to nonsensical results. For example, having two independent error rates in the two-pair case allows EU to explain any observed pattern proportions (leaving zero degrees of freedom), and results in negative degrees of freedom for more general theories. A middle ground is to make error rates depend on some feature of the choice—e.g., how "close" the gambles are (in expected value or in a Euclidean metric applied to the triangle diagram), or how costly an error is. The main obstacle to doing this well is to develop a theory of decision cost. It is inappropriate to assume that an error is less likely in a pair of choices with high EV, say, unless EV is the theory being tested. Then the problem of determining decision cost becomes recursive: The cost of an error in a particular choice pair depends on the theory being tested. We don't think that more complex theories (beyond EU, say) will generate strong restrictions on error rates across choices, but it would be useful to try.

Third, theorists have developed alternatives to EU emphasizing structural explanations: EU axioms are weakened to encompass a broader set of behaviors (additional patterns in this case). Our approach reflects this emphasis by testing pattern-based explanations of choice with the simple, restrictive assumption of independent and equal errors. Again, another approach is to combine a theory of decision cost with less expansive structural explanations.[11] Yet another path

[11] Nat Wilcox commented that more sophisticated error explanations may generate related error rates that differ for each pairwise choice. Having two independent error rates in the two-pair case does generate nonsensical results, but two dependent error rates may be sensible if justified by a more sophisticated theory of errors. The hard work of constructing such a theory remains.

is to test specific parametric forms of theories and allow choices to be stochastic. We think parametric estimation of this sort, with associated error theories, is an important direction for further research; Camerer and Ho (in press) and Hey and Orme (1993) are a start. Our approach, and the more sharply focused parametric approach, are complementary. We test theories in their fullest generality; if a theory is rejected using our method, we can safely abandon it. The parametric approach, on the other hand, could show that a theory which passes our tests is still difficult to specify parsimoniously. For example, we test a very general form of prospect theory, which fits reasonably well. Further tests are needed to establish whether there is a simple family of probability weighting functions—which are central in prospect theory—that also fit well (there appears to be; see Tversky and Kahneman (1992), Camerer and Ho (in press).

Fourth, we assume errors are independent because we find no form of dependence persuasive. If we truly interpret the errors as "error"—like trembles in noncooperative games—then dependence seems illogical. One can imagine alternative assumptions. For example, a thoughtful referee suggested an example in which one theory predicts a pattern *RR* and another predicts *RS*, and the data consist of 1/3 choices of each *SS*, *RR*, and *RS*. Under our approach the *RS* theory predicts best, since it can explain the *SS* and *RR* choices as only one error away from *RS*. *RR* theory predicts poorly because *RS* patterns are one error away but *SS* patterns are two errors away. We think *RS should* be considered better. To rank the *RR* and *RS* theories as equally good is to assume that *RS* and *SS* deviations from the *RR* pattern are equally likely, which forces us to think of an error as the choice of an *entire pattern* against true preference (rather than a particular choice against preference). This route takes us back where many studies started—by simply adding up the fraction of unpredicted patterns. Another argument against this route is *reductio ad absurdum*: This path requires us to think that a person who actually prefers $RR \ldots RR$ (n times) is equally likely to err by choosing $RR \ldots RS$ as by choosing $SS \ldots SS$. That seems to invoke an unnatural theory of errors.

Table III shows maximum-likelihood estimates for several theories. For example, the estimated fanning out proportions are .320, .072, .327, and .281; the estimated error rate is .073. Comparing these estimates with unrestricted proportion estimates gives a log likelihood chi-squared statistic (X^2) testing the goodness-of-fit of the fanning out hypothesis.[12] For fanning out, $X^2 = 1.9$ with 3 degrees of freedom[13] ($p = .588$), so we cannot reject the restriction on true pattern proportions imposed by fanning out.

The maximum-likelihood test is more powerful than the z test described above; theories which survive the z test may not survive the maximum-likeli-

[12] That is, the chi-squared statistic tests the hypothesis that the underlying proportions of people with preferences for *SSR*, *SRS*, *SRR*, and *RSR* are zero. Of course, the predicted proportions of people exhibiting these four patterns of preference will still be positive because of the error rate.

[13] The number of degrees of freedom for a theory is the number of patterns minus the number of linearly independent parameters minus one (so that expected frequencies under the maximum likelihood estimates add to the total sample size).

TABLE III

BATTALIO, KAGEL, AND JIRANYAKUL: REAL LOSSES, SERIES 1

Pattern 123[a]	Observed Frequency	EU	Fan Out	Fan In	MF	RD-cave gIE	RD-cave fDE	RD-vex gDE	RD-vex fIE	PT
SSS	7	.435	.320	.435	.283	.295	.339	.349	.309	
SSR	1			0	.026	.023	0	.003		
SRS	1							0	.003	
SRR	1			0	.029		0	.012	.027	
RSS	3		.072		.092	.081	.164		.082	
RSR	0				0	0	0			
RRS	8		.327		.315	.322		.372	.319	1
RRR	7	.565	.281	.565	.255	.279	.497	.264	.260	0
	$n = 28$									
Error Rate		.209	.073	.209	.038	.059	.183	.095	.056	.357
Chi-squared Statistic		15.8	1.9	15.8	1.0	1.6	14.4	2.8	1.5	17.2
Degrees of Freedom		5	3	3	0	1	1	1	1	5
P Value		.007	.588	.001	0	.200	$1.5E-4$	.093	.220	.004
Posterior Odds for EU[b]			0.03	28.0	2.47	0.65	380	1.17	0.61	2.02

[a] Outcomes: −$20, −$12, $0. Probabilities: $S1(.6, .4, 0)$, $R1(.84, 0, .16)$; $S2(0, 1, 0)$, $R2(.6, 0, .4)$; $S3(0, .2, .8)$, $R3(.12, 0, .88)$.
[b] Posterior odds for EU against each model under minimal prior information.

hood test. For example, EU performs well compared to a random-choice benchmark: It allows 25% of the possible patterns, and accounts for 50% of actual patterns chosen for a z statistic of 3.1. Table III shows, however, that EU cannot explain the systematic variation in its inconsistent patterns. EU predicts that *SRR*, *RSR*, and *RRS* will all be chosen equally often (the maximum likelihood estimates give an expected frequency for each of three patterns of 2.49), but the observed frequencies for the three patterns are 1, 0, and 8. While EU does well by the z test, the chi-squared test shows that EU is unable to account for the variation in the inconsistent patterns ($p = .007$).

The maximum-likelihood test gives two indications of predictive adequacy that aid in diagnosing why some theories fit the data poorly. First, a poor theory must invoke a high error rate to explain frequent choice of patterns it did not allow. For example, prospect theory allows true patterns of *RRS* and *RRR* but many subjects chose *SSS*. A high error rate (.357) is needed to explain why so many subjects chose *SSS* (since the theory interprets the *SSS* choices as two errors by people who truly prefer *RRS*, or three errors by those who prefer *RRR*). Direct estimates of error rates, derived by having subjects make the same choice twice without realizing it, suggest a natural rate of 15–25% (Starmer and Sugden (1989), Camerer (1989), and unpublished data collected by Harless; cf. Battalio, Kagel and Jiranyakul (1990, fn. 13)). Error rates much higher than the natural rates, like the .357 estimated for prospect theory, indicate a poor fit. Error rates which are much lower, like the .038 estimated from mixed fanning, indicate overfitting (i.e., using too many allowed patterns, rather than natural error, to explain the distribution of choices).

Second, poorly fitting theories have patterns for which the maximum-likelihood estimate of the true proportion is zero. Such a theory sacrifices parsimony with no increase in accuracy. For example, the mixed fan hypothesis allows seven of eight patterns in Table III. Coupled with an error rate, that should be enough free proportion parameters to exactly fit the observed choices ($X^2 = 0$). But it isn't: One pattern probability, $p(RSR)$, is estimated to be zero (its unconstrained maximum-likelihood value was negative); the chi-squared statistic for mixed-fan is therefore 1.0 and, with no degrees of freedom, its p-value is zero.

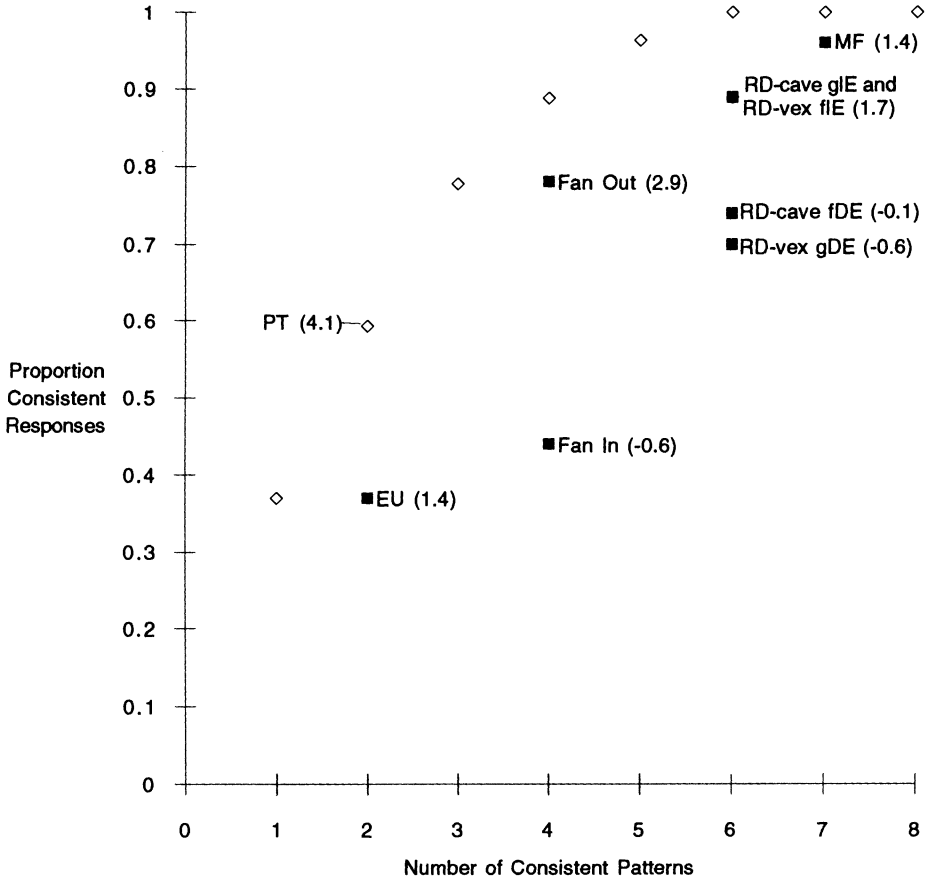
The chi-squared test makes the fit-parsimony tradeoff explicit, but does not resolve the problem of picking the best theory. Theories with more patterns obviously ought to fit better, but how much better must the fit be to justify additional proportion parameters? A formal way to evaluate theories with different numbers of parameters is the "minimal prior information" posterior odds criterion. Klein and Brown (1984) show that when prior information in an experiment is minimized (the expected gain in information from the experiment is made much larger than the information in the prior) the Bayesian posterior odds criterion for Model 1 against Model 2 is $[n^{-(K_1-K_2)/2}]$ [Maximized Likelihood under Model 1/Maximized Likelihood under Model 2], where n is the sample size and K_1 and K_2 are the number of free parameters in the two models. The second term measures the comparative fit of the two models, while the first term adjusts the fit for the difference in dimension of the two models and the sample size.[14]

The posterior odds for EU against each other theory are shown at the bottom of Table III. Fanning out generates the smallest posterior odds (0.03) for EU, providing strong evidence against EU.

Posterior odds is one of several criteria for selecting between models of different parsimony. In Section 3 we discuss some other criteria for model selection; most of them treat unparsimonious theories less harshly than posterior odds do. Posterior odds and other model selection criteria also neglect the estimated error rate, which could (in principle) be traded off against fit and parsimony. We include posterior odds simply as a suggestion for how fit and parsimony might be weighed and to impose consistency on those tradeoffs.

Table III showed data in which subjects actually suffered a loss (from a stake of money given to them initially). Figure 3 and Table IV show results from hypothetical choices over the same set of gambles. (In experiments with hypo-

[14] Another formal way to evaluate some of the theories uses nested hypothesis tests. For example, the EU restriction on pattern proportions is nested within the fanning out restriction. Where hypotheses are nested, the reader can easily undertake such an hypothesis test by subtracting the goodness-of-fit chi-squared statistics in the tables. For example, in Table III the chi-squared statistic testing the EU restrictions on pattern proportions against the fanning out restrictions is $(15.8 - 1.9) = 13.9$ with $(5 - 3) = 2$ degrees of freedom; the nested hypothesis test rejects the EU restrictions ($p = 0.001$). We do not report the nested hypothesis tests because they add little information beyond the goodness-of-fit statistics and because there are many cases where PT and EU are nonnested so the test cannot be applied.



Z statistics (in parentheses) test each theory against the random choice null hypothesis.

FIGURE 3.—Battalio, Kagel, and Jiranyakul: Hypothetical losses, Series 1.

thetical choices subjects were instructed to choose as if one of their choices would be played out for real payoffs.) Comparing the figures provides a glimpse of how motivating subjects, by playing one of the gambles they chose, affects their choices. Figure 2 (real) and Figure 3 (hypothetical) look similar. Prospect theory, fanning out, and RD-vex-f IE are undominated in Figure 2; prospect theory, fanning out, RD-vex-f IE, and mixed fanning are undominated in Figure 3.

The maximum-likelihood error rate analyses reported in Tables III and IV show some subtle differences which are hidden by the figures. Compared to data with hypothetical losses (Table IV), the data for real losses (in Table III) have lower error rates (except for PT). It appears that paying subjects reduces variance (Smith and Walker (1993)). But paying subjects does not increase their adherence to EU. Instead, the lower variance in the real-loss data implies that EU is rejected with real data ($p = .007$), but fits better with hypothetical data

TABLE IV

BATTALIO, KAGEL, AND JIRANYAKUL: HYPOTHETICAL LOSSES, SERIES 1

Pattern 123[a]	Observed Frequency	EU	Fan Out	Fan In	MF	RD-cave gIE	RD-cave fDE	RD-vex gDE	RD-vex fIE	PT
SSS	0	0	0	0	0	0	0	0	0	
SSR	0			0	0	0	0	0		
SRS	1							0	0	
SRR	2			0	.063		0	0	.014	
RSS	5		.207		.186	.190	.273		.209	
RSR	3				.088	.048	0			
RRS	6		.225		.247	.238		.406	.226	.406
RRR	10	1	.568	1	.416	.524	.727	.594	.551	.594
	$n = 27$									
Error Rate		.284	.130	.284	.055	.111	.180	.204	.123	.204
Chi-squared Statistic		11.7	2.5	11.7	1.6	2.3	5.6	6.7	2.5	6.7
Degrees of Freedom		5	3	3	0	1	1	1	1	5
P Value		.039	.476	.009	0	.131	.018	.010	.116	.243
Posterior Odds for EU			0.27	27.0	24.3	6.6	34.9	60.6	7.3	0.08

[a] Outcomes: −$20, −$12, $0. Probabilities: $S1(.6, .4, 0)$, $R1(.84, 0, .16)$; $S2(0, 1, 0)$, $R2(.6, 0, .4)$; $S3(0, .2, .8)$, $R3(.12, 0, .88)$.

($p = .039$). (The z statistics paint the opposite, misleading, picture: EU does better with real losses, $z = 3.1$, then hypotheticals, $z = 1.4$.)

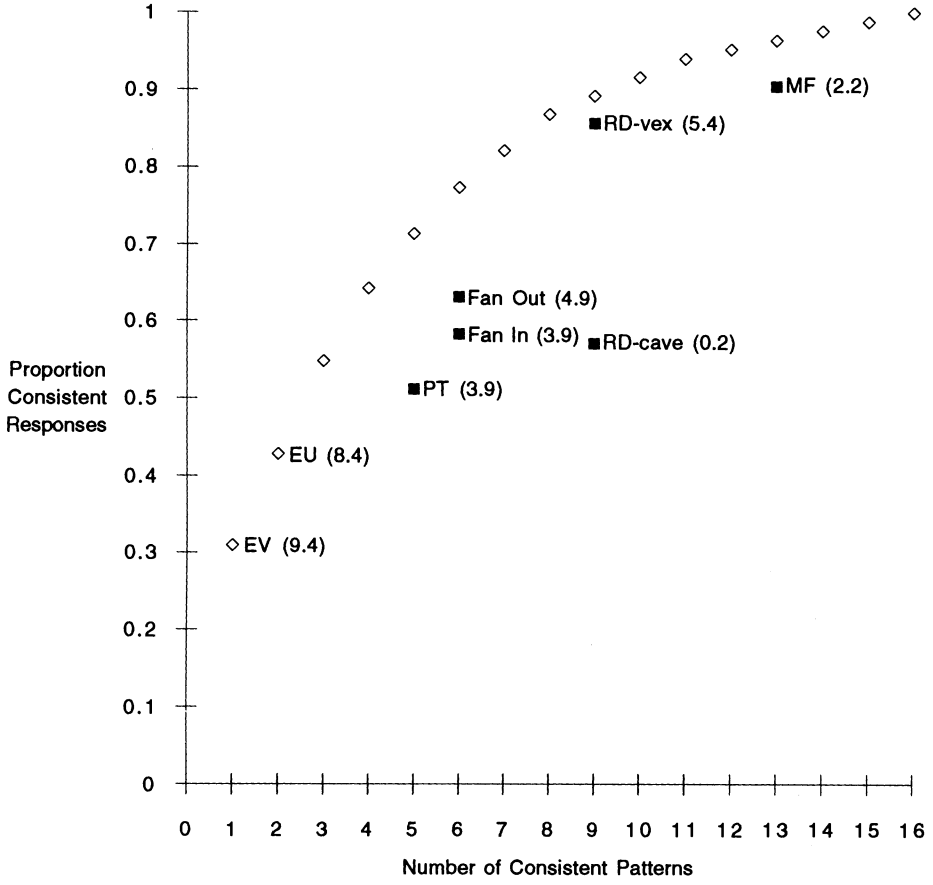
In Figures 2 and 3 prospect theory dominates EU; both theories allow two patterns but prospect theory picks out the two most highly occupied patterns for both real losses and hypothetical losses. The error rate analyses show that prospect theory generates an excellent fit for the hypothetical losses (Table IV), but generates a poor fit—marginally worse than EU—for real losses (Table III). Risk preference for losses makes PT as parsimonious as EU, but that parsimony comes at too high a price for real losses: the highly occupied *SSS* pattern is excluded.

2. OTHER CHOICE STUDIES

In this section we review the results of three other studies. We highlight the distinctive features of each study and draw some conclusions. The results of these and several other studies are formally aggregated in Section 3.

Harless

Harless (1992) examined choices over real gain and real loss lotteries (one lottery was played with real payoffs) in common consequence lottery pairs just inside the triangle boundary. Some people have suggested that systematic deviations from EU disappear in the triangle interior (Conlisk (1989), Camerer (1992)). If it is true, this fact is important. A gamble on the boundary has some outcomes which have zero probability. Moving off the boundary into the interior means that an outcome which had zero probability now has positive probability. Therefore, the disappearance of deviations as one moves from the boundary to



Z statistics (in parentheses) test each theory against the random choice null hypothesis.

FIGURE 4.—Harless: Real gains from unit triangle interior.

the interior suggests the source of the deviations may be nonlinear weighting of low probabilities (cf. Neilson (1992b)).

The conclusion appears to be overstated, at least for gains. The results are shown in Figure 4 and Table V for gains, and in Figure 5 and Table VI for losses. The tables and figures show the responses of Harless's original subjects plus the responses of 38 more subjects.[15]

Figure 4 shows the data frontier for gains. The figure is useful for screening out RD-cave ($z = 0.2$) and fanning in (which is dominated by fanning out), but does not help distinguish among the other theories. The chi-squared error rate analysis in Table V rules out several theories which pass the z test—for

[15] We recruited undergraduates from Wharton as additional subjects (using exactly the same procedures as in the original study) to bring the sample size to a level appropriate for the chi-squared test. We also gathered additional responses to augment the Chew and Waller (1986) data set. The test for the explanatory power of the models over the entire distribution of responses requires a larger sample size than the test of models' performance compared to random choice.

TABLE V

HARLESS: REAL GAINS FROM UNIT TRIANGLE INTERIOR

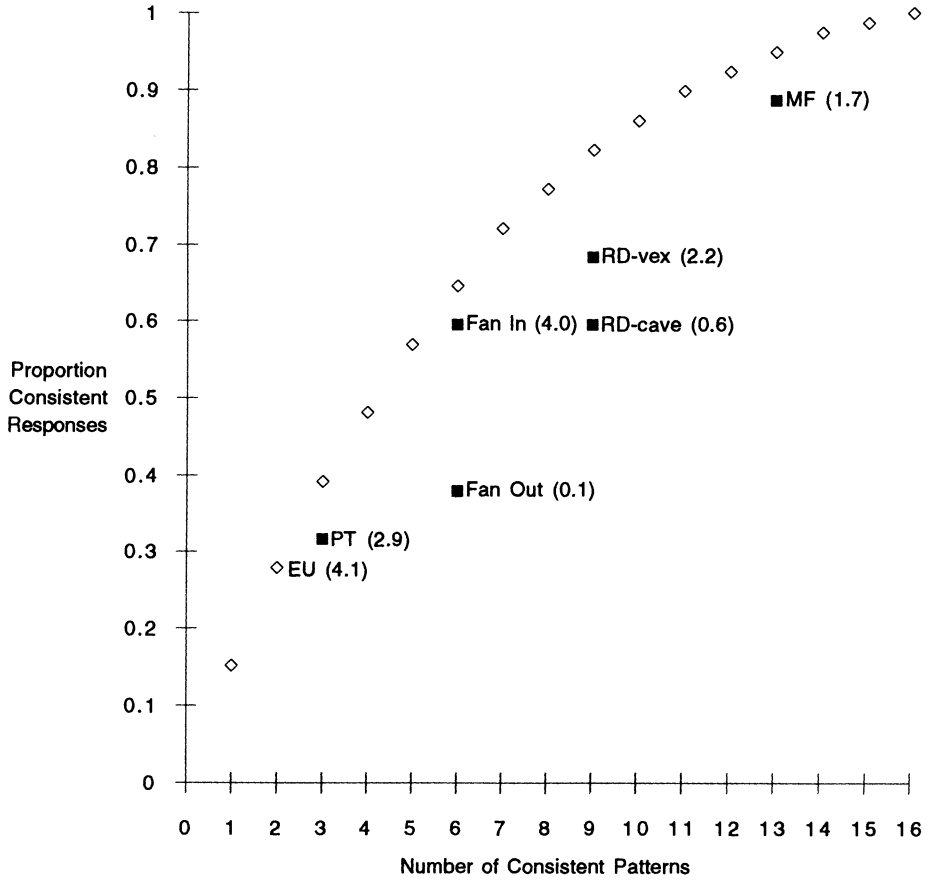
Pattern 1357[a]	Observed Frequency	EV	EU	Fan Out	Fan In	MF	RD-cave	RD-vex	PT
SSSS	10		.256	.213	.252	.174	.252	.187	.252
SSSR	2				0	.009	0		0
SSRS	2						0		
SSRR	4				.010	.053	.010		.010
SRSS	2							0	
SRSR	1				0	0		0	
SRRS	4							.050	
SRRR	6				0	.083		.094	
RSSS	1			0		0	0		
RSSR	1					0	0		
RSRS	1			0		.002	0		
RSRR	1					0	0		0
RRSS	10			.173		.147		.150	
RRSR	8					.080		.071	
RRRS	5			0		.057		.019	
RRRR	26	1	.744	.614	.738	.395	.738	.429	.738
	$n = 84$								
Error Rate		.366	.219	.166	.216	.092	.216	.107	.216
Chi-squared Statistic		59.9	29.2	15.8	29.2	7.1	29.2	8.8	29.2
Degrees of Freedom		14	13	9	9	2	6	6	10
P Value		$1.2E-7$	.006	.071	.001	.029	$5.7E-5$	.186	.001
Posterior Odds for EU		$5.1E+5$		8.6	6,902	$5.9E+5$	$5.3E+6$	200	753

[a] Outcomes: \$0, \$3, \$6. Probabilities: *S*1(.84, .14, .02), *R*1(.89, .01, .10); *S*3(.04, .94, .02), *R*3(.09, .81, .10); *S*5(.44, .14, .42), *R*5(.49, .01, .5); *S*7(.04, .14, .82), *R*7(.09, .01, .9).

example, EV has the highest z statistic but has the lowest chi-squared p value. The conjecture that EU violations disappear in the interior appears to be false, since the chi-squared test gives a p value of .006. Nevertheless, no other theory accounts for the distribution of non-EU choices parsimoniously. Fanning out, mixed fan, and RD-vex have higher p values than EU, but they waste degrees of freedom on sparsely occupied patterns.

The posterior odds ratios favor EU over all competitors, showing that while EU *is* systematically violated, its competitors are no more accurate (adjusting for the number of patterns they allow). However, EU has a larger error rate than more general theories; if we could trade off error rates with fit and parsimony, other theories might look better. For example, RD-vex has a higher p value than EU (.186 versus .006) and a lower error rate (.107 versus .219), but the posterior odds of EU against RD-vex are 200-to-1. Forcing EU to have a lower error rate would shift the odds toward RD-vex.[16] Furthermore, the poor

[16] The error rate is always at least as large for EU than for more general theories which include EU, but the posterior odds criterion does not penalize EU for its high error rate. Dave Grether suggested a way to correct this bias, by computing posterior odds after restricting the error rate to be the same for all theories. However, there is no obviously correct way to choose a single rate, or estimate one from the data. The reader should keep in mind that the posterior odds we report put EU in the best possible light.



Z statistics (in parentheses) test each theory against the random choice null hypothesis.

FIGURE 5.—Harless: Real losses from unit triangle interior.

absolute fit of EU ($p = .006$) means there is room for improvement: A theory which restricts fanning out, allowing pattern *RRSS* but ruling out the other non-EU patterns, would lead to strong evidence against EU (posterior odds of 0.01 for EU).

Figure 5 and Table VI give results for gambles over small losses. Again the z statistic can mislead: EU has a higher z statistic for gains ($z = 8.4$) than losses ($z = 4.1$), but the chi-squared p values are reversed ($p = .006$ for gains, $p = .133$ for losses). In both cases, EU beats all competitors by the posterior odds ratio.

In the BKJ study prospect theory poorly fit the real loss data from the triangle boundary. Here prospect theory poorly fits loss data from the triangle interior (Table VI). In both loss studies the R gambles are mean-preserving spreads of the S gambles. For border gambles in BKJ, risk preference for losses makes PT as parsimonious as EU. For interior gambles in the Harless study, risk preference for losses makes PT nearly as parsimonious as EU. In both cases prospect

TABLE VI

HARLESS: REAL LOSSES FROM UNIT TRIANGLE INTERIOR

Pattern 1357[a]	Observed Frequency	EU	Fan Out	Fan In	Mixed Fan	RD-cave	RD-vex	PT
SSSS	10	.430	.430	.229	.201	.232	.297	
SSSR	9			.154	.138	.268		
SSRS	2					0		
SSRR	3			0	0	0		
SRSS	4						0	
SRSR	7			.167	.111		.251	.650
SRRS	3						0	
SRRR	6			.025	.103		.017	0
RSSS	1		0		0	0		
RSSR	2				0	0		
RSRS	1		0		0	0		
RSRR	7				.101	0		
RRSS	4		0		.063		.007	
RRSR	6				.057		.004	
RRRS	2		0		0		0	
RRRR	12	.570	.570	.425	.226	.500	.424	.350
	$n = 79$							
Error Rate		.281	.281	.222	.141	.248	.236	.362
Chi-squared Statistic		18.7	18.7	7.4	3.2	11.3	10.2	22.8
Degrees of Freedom		13	9	9	2	6	6	12
P Value		.133	.028	.592	.204	.080	.117	.030
Posterior Odds for EU			6,241	22.5	$1.2E+7$	$1.1E+5$	$6.2E+4$	68.2

[a] Outcomes: −$4, −$2, $0. Probabilities: *S*1(.8, .18, .02), *R*1(.88, .02, .1); *S*3(.02, .96, .02), *R*3(.1, .8, .1); *S*5(.41, .18, .41), *R*5(.49, .02, .49); *S*7(.02, .18, .80), *R*7(.10, .02, .88).

theory's fit is worse than that of EU because the common pattern *SSSS* is excluded.

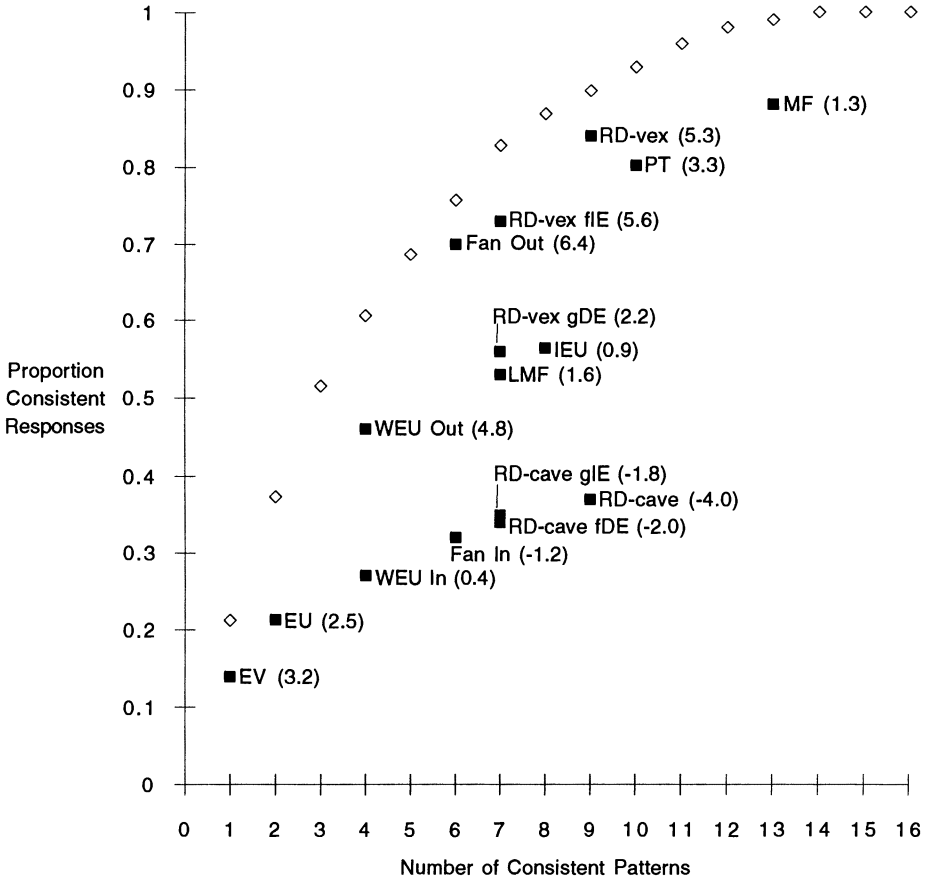
For losses, note that fanning in is the only challenger to EU as a parsimonious theory that also achieves reasonable fit—in sharp contrast to the fanning out for gains in the triangle interior (Table V). Fanning in for losses inside the triangle is also prevalent in our reanalysis of data from Camerer (1992). Mixed fan can account for both fanning out for gains and fanning in for losses inside the triangle, but it fits poorly because it uses too many free parameters to explain patterns that are occupied only because of random error.[17]

[17] In the Harless and BKJ experiments subjects were allowed to respond that they were indifferent between two lotteries. For example, in Harless's study using real gain lotteries, in addition to the 84 subjects in Table V, there were two subjects that indicated they were indifferent between lotteries *S*1 and *R*1. Letting *I* represent indifference, their responses were *IRSR* and *ISRR*. We exclude such responses from our main analysis, but in footnote 26 we summarize how little the results change when indifference responses are included. We assume that the indifference response may belong to one of several patterns and assign indifference responses to maximize the likelihood function for each theory. For example, for each theory an *IRSR* response may be assigned to the *SRSR* or *RRSR* pattern, whichever yields a higher likelihood for the theory.

Chew and Waller

The Chew and Waller (1986) study combines three common consequence choices with a common ratio lottery choice, allowing a test of whether indifference curves are linear. Figure 6 and Table VII contain the responses to the Chew and Waller hypothetical small gain choices on the triangle boundary (their context 1A) for the 56 subjects in their study and 43 new subjects we recruited. Figure 6 suggests that violations of EU appear to be due to fanning out or convex indifference curves (RD-vex): theories which lack these features are generally dominated by theories which have them.

The figure does not show whether the assumption of parallel indifference curves (EU) should be sacrificed for linear indifference curves that fan out (WEU-out), general fanning out, or RD-vex, since all those theories have high z statistics. The chi-squared test in Table VII provides a definite answer. Fanning out fits quite well ($p = .087$), with fewer patterns than other theories with



Z statistics (in parentheses) test each theory against the random choice null hypothesis.

FIGURE 6.—Chew and Waller: Hypothetical moderate gains, Context 1A.

TABLE VII

CHEW AND WALLER: HYPOTHETICAL MODERATE GAINS, CONTEXT 1A

Patterns OILH[a]	Observed Frequency	EV	EU	WEU Out	WEU In	LMF	IEU	Fan Out	Fan In	MF	RD-cave	RD-cave *g*IE	RD-cave *f*DE	RD-vex	RD-vex *g*DE	RD-vex *f*IE	PT
SSSS	7		.358	.173	.358	.173	.173	.152	.358	.111	.090	.090	.090	.130	.127	.134	.119
SSSR	3				0	0	0		0	.018	0	0	0				.033
SSRS	9			.146		.146	.146	0		.056	.415	.415	.415				.025
SSRR	0					0	0			0	0	0	0				0
SRSS	8													.039	.246	.067	
SRSR	4								0	.057				.054	.009		
SRRS	21							.480		.363				.403		.418	.434
SRRR	7									.016				0			.018
RSSS	1										0						
RSSR	1								0	.002	0		0				
RSRS	2							0		0	0	0					0
RSRR	0									0	0	0	0				0
RRSS	3						0							0	0	0	
RRSR	3				0	0	0		0	.016				.011	0	.025	
RRRS	16			.494		.494	.494	.161		.192				.182	.452	.171	.176
RRRR	14	1	.642	.187	.642	.187	.187	.207	.642	.169	.495	.495	.495	.181	.166	.185	.195
	$n = 99$																
Error Rate		.452	.339	.243	.339	.243	.243	.178	.344	.121	.301	.301	.301	.143	.234	.157	.157
Chi-squa. ed Statistic		86.9	82.0	49.9	82.0	49.9	49.9	15.2	82.0	10.3	71.1	71.1	71.1	11.2	44.5	13.4	14.2
Degrees of Freedom		14	13	11	11	8	7	9	9	2	6	8	8	6	8	8	5
P Value		$1E-12$	$5E-12$	$6.4E-7$	$6.1E-13$	$4.2E-8$	$1.5E-8$	.087	$6E-14$	.006	$2E-13$	$3E-12$	$3E-12$	.084	$5E-7$	.098	.015
Posterior Odds for EU		1.2		$1.1E-5$	99.0	0.01	0.11	$3.0E-11$	9,801	$2.6E-5$	$4.1E+4$	411	411	$4.0E-9$	$7.2E-4$	$1.3E-10$	$1.8E-7$

[a] Outcomes: \$0, \$40, \$100. Probabilities: *SO*(0, 1, 0), *RO*(.5, 0, .5); *SI*(0, 1, 0), *RI*(.05, .9, .05); *SL*(.9, .1, 0), *RL*(.95, 0, .05); *SH*(0, .1, .9), *RH*(.05, 0, .95).

comparably good fits, and has the lowest posterior odds ratio ($3.0E-11$, the strongest evidence against EU).[18]

All the theories that assume linear indifference curves—EV, EU, WEU, linear mixed fan (LMF), and implicit EU—fit poorly by the chi-squared test.[19] Chew and Waller (1986) concluded that WEU-out was superior to EU because its z statistic was higher. In our analysis (including 43 new subjects), WEU-out does provide a much better fit than EU using the chi-squared test and posterior odds ratio, but those tests also show that several theories with nonlinear indifference curves are even better than WEU-out.

Table VII shows another way in which the maximum-likelihood error-rate analysis is more informative than simply counting pattern frequencies. Nine subjects chose pattern *SSRS*, many more than expected under random choice. Fanning out allows that pattern, but its maximum likelihood estimate for the proportion of subjects who truly prefer *SSRS* is zero. Introducing the parameter $p(SSRS)$ to the model that already had consistent patterns *SSSS*, *SRRS*, *RRRS*, and *RRRR* did not improve the fit even though *SSRS* choices were common. The estimated proportion of subjects with true *SSRS* pattern preference is low because raising the estimate increases the expected frequencies of neighboring patterns (*RSRS*, *SRRS*, *SSSS*, and *SSRR*), but those expected frequencies already exceed observed frequencies. Fanning out could afford to exclude *SSRS*, even though it is chosen often, because it is fed by errors from the *SSSS* and *SRRS* patterns and therefore has a high expected frequency (6.6) even if *SSRS* is excluded.

Sopher and Gigliotti

Sopher and Gigliotti (1990) gathered responses to the Allais Paradox common consequence pairs and three other common consequence pairs from the triangle boundary (Table VIII). They also gathered responses to comparable choices in the triangle interior (Table IX). When lotteries lie in the interior, the EV pattern (*RRRRR*) is chosen much more often. The chi-squared test in Table VIII shows that EU has a terrible fit for lotteries on the boundary of the triangle ($p = 1.5E-25$). EU fits substantially better for interior lotteries, but the p value is still low ($3.7E-10$). By the posterior odds ratios EU is worse than many theories for boundary lotteries (Table VIII) but better than every alternative except prospect theory for interior lotteries (Table IX). (In Harless's data using interior lotteries, Tables V and VI, EU is better than every other theory by the posterior odds ratio.) The success of prospect theory stems from increased precision for interior gambles (allowing only 7 of 32 patterns) com-

[18] Becker and Sarin (1987) note that their lottery dependent utility theory accounts for 98–99% of the choices by allowing 14 of the 16 possible patterns. But one can account for the distribution of responses with far fewer patterns; the posterior odds for fanning out against lottery dependent EU are $2.27E+06$.

[19] Camerer and Ho (in press) reach a similar conclusion from a review of ten studies testing the betweenness axiom (which creates linear indifference curves).

TABLE VIII

SOPHER AND GIGLIOTTI: COMMON CONSEQUENCE HYPOTHETICAL LARGE GAINS ON UNIT TRIANGLE BOUNDARY

Pattern 12345[a]	Observed Frequency	EV	EU	WEU Out	WEU In	LMF	IEU	Fan Out	Fan In	MF	RD-cave	RD-vex	PT
SSSSS	4		.114	0	.114	0	0	0	.114	.023	.016	.088	.024
SSSSR	2												
SSSRS	0						0				0		0
SSSRR	1										0		0
SSRSS	1				0	0	0		0	.001	0		0
SSRSR	1								0	0			
SSRRS	1						0				0		0
SSRRR	1								0	0	0		0
SRSSS	8			.218		.116	.116	.191		.053	.104		.064
SRSSR	6												
SRSRS	6			.116		.111	.111	0		.022	.002		.004
SRSRR	25							.334		.176	.221		.206
SRRSS	8					.150	.150			.043	.095		.099
SRRSR	18									.125			
SRRRS	7					.002	.002			.038	0		0
SRRRR	25									.141	.251		.256
RSSSS	0											0	
RSSSR	1						0					0	
RSSRS	0												
RSSRR	1						0					0	
RSRSS	2								0	.011		0	
RSRSR	1				0	0	0		0	.007		0	
RSRRS	0												
RSRRR	1				0	0	0		0	0		0	
RRSSS	5							0		.027		.043	.028
RRSSR	0						0					0	
RRSRS	4							0		.018			.013
RRSRR	11			.029		.095	.095	0		.051		.135	.023
RRRSS	2									.007		0	0
RRRSR	5					0	0			.009		0	
RRRRS	2									0			0
RRRRR	37	1	.886	.637	.886	.526	.526	.475	.886	.248	.311	.734	.283
	$n = 186$												
Error Rate		.343	.299	.237	.299	.212	.212	.198	.299	.060	.147	.280	.129
Chi-Square Statistic		200.2	189.9	144.8	189.9	135.5	135.5	97.4	189.9	15.4	53.9	186.1	49.1
Degrees of Freedom		30	29	26	26	20	15	23	23	10	18	18	13
P Value		$4.6E-27$	$1.5E-25$	$1.9E-18$	$7.4E-27$	$3.6E-19$	$1.8E-21$	$4.0E-11$	$3.2E-28$	.118	$1.9E-05$	$5.9E-30$	$4.3E-06$
Posterior Odds for EU		12.5		$4.1E-7$	2,537	0.02	$1.2E+4$	$5.3E-14$	$6.4E+6$	$4.6E-17$	$9.1E-18$	$5E+11$	$3.8E-13$

[a] Outcomes: \$0, \$1M, \$5M. Probabilities: *S*1(0, 1, 0), *R*1(.01, .89, .1); *S*2(.89, .11, 0), *R*2(.9, 0, .1); *S*3(0, .11, .89), *R*3(.01, 0, .99); *S*4(.79, .11, .1), *R*4(.8, 0, .2); *S*5(.01, .89, .1), *R*5(.02, .78, .2).

TABLE IX

SOPHER AND GIGLIOTTI: COMMON CONSEQUENCE HYPOTHETICAL LARGE GAINS FROM UNIT TRIANGLE INTERIOR

Pattern 12345[a]	Observed Frequency	EV	EU	WEU Out	WEU In	LMF	IEU	Fan Out	Fan In	MF	RD-cave	RD-vex	PT
SSSSS	17		.235	.231	.235	.189	.175	.231	.179	.152	.175	.195	.175
SSSSR	1												
SSSRS	0						0				0		
SSSRR	2										.006		
SSRSS	1				0	0	0		0	0	0		.016
SSRSR	0								0	0			
SSRRS	9						.084				.072		.064
SSRRR	6								.102	.056	.019		.023
SRSSS	2			0		0	0	0		0	0		
SRSSR	0												
SRSRS	2			.010		0	0	.010		0	.002		
SRSRR	2							0		0	0		
SRRSS	4					.012	.024			.019	.028		
SRRSR	3									0			
SRRRS	8					.136	.066			.084	.039		.058
SRRRR	21									.113	.122		.122
RSSSS	8											.012	
RSSSR	0						0					0	
RSSRS	0												
RSSRR	1						0					0	
RSRSS	5								.055	.059		.050	
RSRSR	2				0	.011	.014		0	0		0	
RSRRS	3												
RSRRR	7				0	.016	0		0	.004		0	
RRSSS	0							0		0		0	
RRSSR	3						.006					0	
RRSRS	0							0		0			
RRSRR	5			0		0	0	0		0		0	
RRRSS	2									0		0	
RRRSR	5					0	0			.004		0	
RRRRS	4									0			
RRRRR	61	1	.752	.746	.752	.637	.615	.746	.651	.508	.533	.729	.542
	$n = 184$												
Error Rate		.327	.186	.184	.186	.148	.130	.184	.151	.104	.112	.173	.118
Chi-Square Statistic		231.5	102.6	102.3	102.6	77.0	64.1	102.3	81.7	43.8	49.6	95.9	52.6
Degrees of Freedom		30	29	26	26	20	15	23	23	10	18	18	24
P Value		$5.4E-33$	$3.7E-10$	$5.2E-11$	$4.7E-11$	$1.3E-08$	$4.9E-08$	$5.5E-12$	$1.7E-08$	$3.6E-06$	$8.6E-05$	$1.3E-12$	$6.6E-04$
Posterior Odds for EU		$7E+26$		2,189	2,496	$4.3E+4$	$3.1E+7$	$5.5E+6$	184	$5.5E+8$	9.04	$1E+11$	$6.2E-6$

[a] Outcomes: \$0, \$1M, \$5M. Probabilities: *S*1(.01, .98, .01), *R*1(.02, .87, .11); *S*2(.80, .19, .01), *R*2(.81, .08, .11); *S*3(.01, .19, .80), *R*3(.02, .08, .90); *S*4(.70, .19, .11), *R*4(.71, .08, .21); *S*5(.02, .87, .11), *R*5(.03, .76, .21).

pared to boundary gambles (allowing 18 of 32), which serves it well since there are fewer systematic patterns with interior gambles.[20]

There is another interesting difference between interior and boundary results. Fanning out fits better than fanning in for boundary lotteries (Table VIII) and the opposite is true for interior lotteries (Table IX). Neither theory fits well for both sets of gambles.

The data provide little reason to replace the independence assumption in EU with the weaker assumption of betweenness. The theories which assume betweenness have high posterior odds supporting EU. The only exception is WEU-out on the boundary, but in that case several other theories have even better posterior odds than WEU-out.

Perhaps the most striking feature of Tables VIII and IX is the large number of maximum likelihood estimates which equal zero. All the generalizations of EU are guilty. Under the error rate approach, the distribution of responses may be explained with relatively few consistent patterns. EU is too lean (it allows too few patterns to explain the distribution); generalizations of EU are too fat (they predict too many useless patterns or the wrong ones).

3. AGGREGATION OF RESULTS ACROSS STUDIES

Sections 1 and 2 demonstrate that the error rate analysis of choice data is more powerful than statistical tests employed previously. In general, some conclusions drawn from earlier studies are reversed by our analyses—e.g., in some cases a theory with a high z statistic, which predicts much better than a random choice benchmark, cannot account for variation in unpredicted patterns and hence is rejected by the chi-squared test while a theory with a lower z statistic is not rejected by the chi-squared test. And while EU performs relatively better with gambles in the triangle interior (compared to generalizations), by our tests it is easily rejected there too. Some new conclusions appear too—e.g., paying subjects appears to lower the error rate, *increasing* rejection of EU and many other theories rather than inducing conformity to them. Many theories have been proposed to explain violations of EU such as the common consequence and common ratio effect. Our method uses the distribution of responses across combinations of common consequence and common ratio choice problems generating tests with more power showing that some of the

[20] There are some instances, such as the Sopher and Gigliotti data sets, where we do not impose restrictions on a theory's parameters across data sets. We take each set of pairwise choices from a unit triangle as a separate data set. Since subjects were randomly assigned to either the boundary or interior treatment in the Sopher and Gigliotti study, the risk preferences between the two groups should be identical. Hence, one could consider adding the restriction for EU that the proportion of subjects prefering *SSSSS* must be the same for lotteries on the boundary and the interior. The other instance where cross-data set restrictions could be considered involves gain and loss triangles with choice pairs having mean-preserving risk spreads to test prospect theory's reflection effect (Battalio, Kagel, and Jiranyakul (1990), Camerer (1989, 1992)). Since reflection is confirmed in these studies, we suspect that imposing the restrictions might make prospect theory fare better and theories not incorporating reflection, such as EU, fare worse. We do not impose cross-data set restrictions because there is no precise prediction for many of the generalizations of EU.

candidate theories *cannot* explain the distribution of responses while others can. Our method clearly demonstrates that some theories are losers as they are less parsimonious and have a poorer fit than other theories. The posterior odds ratios also provide an unequivocal method for trading off fit and parsimony in comparing EV, EU, and the generalizations. (In contrast, analysis based on the proportion of consistent responses yields conflicting conclusions for each of three measures, difference, inside ratio, and outside ratio.) And the analysis has another important advantage: Since the sum of several independent random variables with chi-squared distributions also has a chi-squared distribution, chi-squared statistics from different experiments may be added to gauge each theory's performance across a variety of subjects, investigators, experimental methods, and so on, as long as the experiments are independent. Aggregation tells us whether deviations from EU are robust across studies.

The independence of studies is hard to evaluate. For our purposes, independence means that conditional on the truth of a particular theory, a sample of results from one experiment did not influence the results from another. Nonindependence could arise because the same subjects are represented by choices in multiple data sets, or because some investigators designed their experiments based on earlier results. The amount of design dependence is difficult to measure empirically. It is surely substantial because virtually all the studies we aggregate used some variation on familiar common consequence, common ratio, betweenness-testing, and framing problems introduced by Allais (1953) and Kahneman and Tversky (1979) and many used a design adapted from Chew and Waller (1986) (proposed initially by Chew and MacCrimmon (1979)). Since we cannot measure the degree of dependence between studies reliably, we offer two caveats to our aggregation analyses (which assume independence): First, violations of independence imply that the p values we compute by adding chi-squared statistics are too low; skeptical readers might adjust them upward to reflect the degree of dependence they think exists. Second, in many cases a single study generates p values low enough to cast severe doubt on one or more theories, so even if studies are perfectly dependent (i.e., are pure replications which differ only by sampling error) the data are sufficient to reject some theories.

With the caveat about our heavy-handed independence assumption in mind, we discuss the result of aggregating different studies. Included in our chi-squared aggregation are the results from the seven data sets presented in Sections 2 and 3 and the results from sixteen more data sets from Battalio, Kagel, and Jiranyakul (1990), Camerer (1989, 1992), Chew and Waller (1986), and Sopher and Gigliotti (1990) presented in an unpublished Appendix (available upon request).[21]

[21] We excluded studies with two-pair patterns. Tests like ours are correct but severely limited when applied in a two-choice case (e.g., Conlisk (1989), Battalio, Kagel, and Jiranyakul (1990), Prelec (1990)). Two pairs (four patterns) do not span a broad range of gambles, and hence are too few to distinguish all the theories we consider. For example, in a two-pair test weighted and implicit EU cannot be distinguished, and mixed fanning allows all four patterns.

TABLE X

PERFORMANCE OF THEORIES USING MEASURES BASED ON PERCENTAGE CONSISTENT RESPONSES

	Sum of Squared Z Statistics	Degrees of Freedom	P Value	Ratio Measure	Difference Measure	Outside Ratio
EV	1,432	23	$9E-289$	4.14	.17	.81
EU	1,183	23	$2E-235$	3.03	.20	.76
WEU-Out	599	23	$8E-112$	1.89	.22	.66
WEU-In	383	23	$6E-67$	1.61	.14	.78
LMF	279	23	$9E-46$	1.31	.15	.53
IEU	36	6	$3E-06$	1.16	.08	.84
Fan Out	547	23	$5E-101$	1.75	.23	.64
Fan In	308	23	$1E-51$	1.45	.13	.80
MF	291	21	$2E-49$	1.25	.16	.41
RD-cave	155	14	$6E-26$	1.27	.12	.78
RD-cave gIE	247	19	$1E-41$	1.33	.16	.67
RD-cave fDE	242	19	$2E-40$	1.31	.14	.72
RD-vex	118	14	$2E-18$	1.22	.12	.73
RD-vex gDE	79	19	$3E-09$	1.13	.06	.86
RD-vex fIE	129	19	$3E-18$	1.21	.10	.77
PT	627	23	$1E-117$	1.73	.24	.57

Consider first the measures of theory performance based on percentage of consistent responses aggregated over all 23 studies (Table X).[22] The z statistics (comparing each theory's percentage of consistent responses to a random choice null hypothesis) are squared (so that they become chi-squared statistics) and summed (keeping the sign when theories generate negative z statistics). The extreme p values show that each model does much better than would be expected if choices were random. This is an important conclusion, but does not tell us whether theories fail to account for variation in excluded patterns and gives no clear way to choose among theories. Also given in Table X for each theory are the ratio measure (the proportion consistent divided by the proportion of consistent patterns), difference measure (proportion consistent minus proportion consistent patterns), and outside ratio (the proportion of inconsistent responses divided by the proportion of inconsistent patterns) weighted by sample size and aggregated over the 23 studies. None of the measures selects EU. The difference measure selects PT. The ratio measure selects the most parsimonious theory (EV). The outside ratio selects the broadest theory which excludes uncommon patterns (mixed fan). Besides having no good reason to choose one measure over the others, the measures of theory performance in Table X are incomplete because they throw away information about the distribution of responses in consistent and inconsistent patterns.

The sum of chi-squared statistics from the maximum likelihood error analysis over all 23 data sets are given in Table XI. All the theories have extremely low p

[22] Even in studies without specific tests of betweenness, WEU-out is tested whenever fanning out is tested; hence, the sum of squared z statistics for WEU-out includes squared z statistics for fanning out from studies that do not test for betweenness. A similar statement holds for WEU-in, rank-dependent theories with elasticity conditions, and LMF.

TABLE XI

SUM OF CHI-SQUARED STATISTICS OVER ALL STUDIES

	Sum of Chi-squared Statistics	Degrees of Freedom	*P* Value
EV	1,289.0	253	$3E-138$
EU	902.5	243	$1E-76$
WEU-Out	611.6	189	$5E-46$
WEU-In*	869.0	189	$1E-87$
LMF*	386.4	96	$2E-36$
IEU*	294.0	50	$3E-36$
Fan Out	520.3	175	$6E-36$
Fan In*	837.7	175	$4E-87$
MF	153.0	58	$2E-10$
RD-cave	289.4	74	$3E-27$
RD-cave *g*IE	327.3	97	$1E-26$
RD-cave *f*DE*	340.9	97	$7E-29$
RD-vex*	428.7	74	$2E-51$
RD-vex *g*DE*	679.1	97	$4E-88$
RD-vex *f*IE*	548.2	97	$4E-64$
PT	400.9	167	$3E-21$

*Theory is dominated by another with a lower chi-squared statistic and at least as many degrees of freedom.

values. Tables X and XI together tell the whole story: each theory performs much better than would be expected if choices were random (Table X), but there is systematic variation in the patterns they don't predict too (Table XI). In Table XI asterisks indicate dominated theories, which fit worse (higher chi-squared statistics) *and* are less parsimonious (fewer degrees of freedom) than some other theory. Half of the theories—all variations of RD-vex, RD-cave *f*DE, fanning in, and all the theories incorporating betweenness except WEU out—are dominated.[23]

Since the results from individual studies in Sections 1 and 2 indicate that theories perform quite differently under different conditions (boundary versus interior, for example), Table XII decomposes the sum of chi-squared statistics by location in the triangle (boundary, interior) and outcomes (large gain, small gain, small loss) of gambles used in the studies. (A decomposition of hypotheti-

[23] We also tested the Quiggin (1982, 1993) anticipated utility (AU) proposal for rank-dependent expected utility, in which $f(.5) = .5$ and $f(p)$ is concave (convex) below (above) .5. This restriction excludes some patterns in the studies reviewed, but even then AU allows many more patterns than other theories do. AU cannot be tested in 8 of the 23 studies because it excludes no patterns at all. (MF excludes some patterns in 21 of 23 studies. EV, EU, PT, and fanning exclude patterns in all 23 studies.) Aggregating over the studies that do test AU, AU generates a log likelihood chi-squared statistic of 147.7 ($df = 55$, $p = 2.0E-10$). In the menu of best theories in Table XIII, AU would be chosen for $m < 1.8$ and MF would be chosen for $1.8 \leqslant m < 2.3$ (but this is not surprising because AU makes fewer predictions, and hence has fewer degrees of freedom, than MF). Further restrictions on the AU $f(p)$ function would make the theory more parsimonious. For example, one could restrict the function to be symmetric around .5 ($f(p) = 1 - f(1-p)$) or require the concave and convex portions to satisfy an elasticity condition. The studies described herein are not efficiently designed to test AU in the Quiggin form; more sharply-designed studies would be useful.

TABLE XII

DECOMPOSITION OF SUM OF CHI-SQUARED STATISTICS

	Boundary of Unit Triangle											
	Large Gains				Small Gains				Small Losses			
	Chi-squared Statistic	Degrees of Freedom	*P* Value	Minimum *m* Value	Chi-squared Statistic	Degrees of Freedom	*P* Value	Minimum *m* Value	Chi-squared Statistic	Degrees of Freedom	*P* Value	Minimum *m* Value
EV	515.8	80	$6E-65$	[7.4]	125.9	34	$2E-12$	[4.9]				
EU	486.0	76	$5E-61$		121.0	33	$6E-12$		49.2	20	$3E-4$	
WEU-Out	295.8	65	$4E-31$	17.3	66.2*	24	$8E-6$	6.1	26.1	13	.016	3.3
WEU-In	482.9*	65	$4E-65$	0.3	121.0*	24	$6E-15$	0.0	43.1*	13	$4E-5$	0.9
LMF	184.5*	43	$2E-19$	9.1	60.9*	11	$6E-9$	2.7	17.8*	3	$5E-4$	1.8
IEU	160.2*	22	$6E-23$	6.0	54.5*	10	$4E-8$	2.9	15.2*	3	.002	2.0
Fan Out	243.7	60	$5E-24$	15.1	28.7	20	.094	7.1	24.4	11	.011	2.8
Fan In	482.4*	60	$3E-67$	0.2	121.0*	20	$2E-16$	0.0	33.2*	11	$5E-4$	1.8
MF	43.7	27	.022	9.0	16.7	2	$2E-4$	3.4	2.6	0	0	2.3
RD-cave	89.3*	26	$7E-9$	7.9	85.2*	9	$1E-14$	1.5	17.1*	3	.001	1.9
RD-cave *g*IE	119.8*	43	$4E-9$	11.1	88.7*	13	$2E-13$	1.6	21.0*	5	.001	1.9
RD-cave *f*DE	113.2	43	$3E-8$	11.3	92.8*	13	$4E-14$	1.4	37.1*	5	$6E-7$	0.8
RD-vex	270.5*	26	$2E-42$	4.3	20.3	9	.016	4.2	5.0	3	.172	2.6
RD-vex *g*DE	463.8*	43	$6E-72$	0.7	67.9*	13	$2E-9$	2.7	14.5*	5	.013	2.3
RD-vex *f*IE	376.0*	43	$1E-54$	3.3	30.3*	13	.004	4.5	9.0	5	.109	2.7
PT	100.5	34	$2E-8$	9.2	41.5	24	.015	8.8	35.7	19	.011	13.5
	Interior of Unit Triangle											
	Large Gains				Small Gains				Small Losses			
	Chi-squared Statistic	Degrees of Freedom	*P* Value	Minimum *m* Value	Chi-squared Statistic	Degrees of Freedom	*P* Value	Minimum *m* Value	Chi-squared Statistic	Degrees of Freedom	*P* Value	Minimum *m* Value
EV	503.4	72	$6E-66$	[80.3]	64.4	24	$1E-5$	[30.7]				
EU	182.3	68	$2E-12$		33.7	23	.070		30.3	23	.141	
WEU-Out	174.1	57	$9E-14$	0.7								
WEU-In	175.4*	57	$6E-14$	0.6								
LMF	111.9*	35	$6E-10$	2.1								
IEU	64.1*	15	$5E-8$	2.2								
Fan Out	174.1*	54	$1E-14$	0.6	19.3	15	.200	1.8	30.1*	15	.012	0.0
Fan In	154.5	54	$1E-11$	2.0	33.6*	15	.004	0.0	13.0	15	.602	2.2
MF	78.7	25	$2E-7$	2.4	7.1	2	.029	1.3	4.2	2	.122	1.2
RD-cave	52.8	20	$9E-5$	2.7	30.9*	8	$1E-4$	0.2	14.1*	8	.079	1.1
RD-vex	104.0*	20	$2E-13$	1.6	11.6	8	.170	1.5	17.3*	8	.027	0.9
PT	110.6	51	$3E-6$	4.2	50.8*	19	$1E-4$	−4.3	61.8*	20	$4E-6$	−10.5

cal vs. real-payoff results showed no interesting, reliable differences other than those noted in Section 1.) For small losses all the studies used mean-preserving risk spreads so the predictions of EV and EU are identical (but see footnote 27).

Some theories are dominated in nearly every category in Table XII (WEU-in, LMF, and IEU). For other theories the decomposition in Table XII reveals areas of strength and weakness of theories. As the individual studies suggested, whether lotteries lie on the boundary (where lottery supports are different) or the interior (where supports are the same) makes a tremendous difference for the performance of EU. On the boundary of the triangle the p values for EU are all quite low, and EU fits only slightly better than EV. In the triangle interior, however, EU manages a miraculous recovery.

The boundary-interior distinction also reveals differences in the performance of the alternative theories. Fanning out dominates fanning in on the boundary, but in the interior fanning in does better for large gains and small losses. Further, since violations of EU are less systematic in the triangle interior, theories which allow the same number of patterns in the triangle interior as on the boundary (all the theories except PT) have a substantial handicap. Nevertheless, there are indications that systematic deviations from EU occur even in the triangle interior (for example, in the Harless and Sopher and Gigliotti data sets) so there may be room for leaner generalizations to improve upon EU even in the triangle interior.

The boundary and interior classification proves quite useful as a diagnostic tool for prospect theory. For small gains and small losses on the boundary of the triangle PT achieves a good fit (except for BKJ real losses in Table III); because the riskier lotteries are (usually) mean preserving risk spreads of the safer lotteries, PT's property of risk aversion for gains and risk preference for losses (except for small probabilities) makes it quite parsimonious. (For small losses PT has just one fewer degree of freedom than EU.) For small gains and small losses in the triangle interior lotteries are again (usually) mean preserving, but PT has an awful fit. The data therefore provide evidence against reflection—risk aversion for gains and risk preference for losses—when lottery supports are the same.[24] Large gains lotteries are not mean preserving, so PT is not very parsimonious; it wastes degrees of freedom on useless patterns and its fit is undistinguished. However, for large gains in the interior PT allows fewer patterns than on the boundary, making it almost as parsimonious as fanning out or WEU; there it fits relatively well.

Table XII also suggests the possibility that curvature of indifference curves may depend on the size of outcomes: RD-cave dominates RD-vex for large gains but is dominated by RD-vex for small gains and small losses. Although the chi-squared statistics differ by orders of magnitude, the effect may be due to

[24] Note that all the experiments which used real-loss payoffs actually deducted losses from an initial stake subjects had been given. If subjects frame these lotteries as choices over net gain outcomes, the reflection effect predicted by prospect theory is diluted. However, reflection is apparent in boundary lotteries with different support.

differences in the locations of the lottery pairs. The possibility that curvature depends on the size of outcomes deserves further investigation.

The decomposition in Table XII is informative, but the central question of the paper remains: Which alternative to EU competes with EU as the best, parsimonious model?

A large statistical literature on model selection criteria gives guidance on trading off fit (chi-squared) and parsimony (degrees of freedom). Many of these criteria are described by the rule "pick the model for which $X^2 - m$(degrees of freedom) is smallest," where the multiplier m penalizes the use of free parameters. The number m is a marginal rate of substitution between chi-squared and degrees of freedom, or the price of precision.

Various authors have proposed values for m. The Klein and Brown minimal information posterior odds criterion corresponds to the Schwarz criterion (1978) $m = \log n$, where n is the sample size. Others have proposed model selection criteria with smaller multipliers, such as the Akaike criterion, $m = 2$ (Akaike (1973)), the local Bayes factor, $m = 3/2$ (Smith and Spiegelhalter (1980)), $m = 1$ (Nelder and Wedderburn (1972)), the posterior Bayes factor, $m = \log(2) = .69$ (Aitkin (1991)), and the simple maximum likelihood criterion, $m = 0$.

The range of values for m indicates substantial controversy over how to trade off fit and parsimony. We are reluctant to recommend a value of m, but we can impose consistency on the reader's selection of m. Return to Table XII. For each generalization of EU we give the minimum value of m that would lead to the selection of EU over that generalization. For example, for large gains on the boundary of the triangle, the minimum m value to select EU over PT is 9.2; if the reward for degrees of freedom is less than 9.2 then PT should be selected over EU. The reader is invited to reflect on his or her preferences for parsimony, ponder the statisticians' advice on appropriate m values, and choose an m.[25]

The reader selecting an m value high enough to choose EU over all the generalizations faces a dilemma, however. In Table XII we also give [in brackets] the value of m that leads to selection of EV over EU. For large gains on the boundary of the triangle, for example, EV is selected over EU for an m value of 7.4 or higher. The EV-EU fit-parsimony comparison provides the most damning evidence against EU: on the triangle boundary a value of m high enough to lead to the selection of EU over all the generalizations necessarily implies the selection of EV over EU. The reader who highly prizes parsimony must choose EV over EU when the supports of the lotteries are different. However, in the triangle interior (where supports are the same) there are reasonable m values for which EU should be selected over the generalizations and also over EV.

Table XIII provides a compact summary of the results of all the studies. For each of the classifications in Table XII, we show the menu of best, parsimonious

[25] On the boundary of the triangle the Schwarz criterion multiplier values are 6.3 for large gains, small gains 5.4, and small losses 4.8. In the triangle interior the Schwarz criterion multiplier values are 6.3 for large gains, small gains 5.5, and small losses 5.5.

TABLE XIII

THE BEST THEORIES FOR VALUES OF THE MULTIPLIER ASSIGNED TO DEGREES OF FREEDOM (m)

Boundary of Unit Triangle (Lotteries with Different Support)

Large Gains		Small Gains		Small Losses	
m Value	Best Theory	m Value	Best Theory	m Value	Best Theory
$m < 4.3$	MF	$m < 0.5$	MF	$m < 0.8$	MF
$4.3 \leqslant m < 7.7$	RD-cave fDE	$0.5 \leqslant m < 0.8$	RD-vex	$0.8 \leqslant m < 1.9$	RD-vex
$7.7 \leqslant m < 10.4$	Fan Out	$0.8 \leqslant m < 3.2$	Fan Out	$1.9 \leqslant m < 13.5$	PT
$10.4 \leqslant m < 14.7$	WEU Out	$3.2 \leqslant m < 6.1$	PT	$13.5 \leqslant m$	EU/EV
$14.7 \leqslant m$	EV	$6.1 \leqslant m$	EV		

Interior of Unit Triangle (Lotteries with Same Support)

Large Gains		Small Gains		Small Losses	
m Value	Best Theory	m Value	Best Theory	m Value	Best Theory
$m < 1.2$	RD-cave	$m < 0.8$	MF	$m < 0.7$	MF
$1.2 \leqslant m < 4.2$	PT	$0.8 \leqslant m < 1.1$	RD-vex	$0.7 \leqslant m < 2.2$	Fan In
$4.2 \leqslant m < 80.3$	EU	$1.1 \leqslant m < 1.8$	Fan Out	$2.2 \leqslant m$	EU/EV
$80.3 \leqslant m$	EV	$1.8 \leqslant m < 30.7$	EU		
		$30.7 \leqslant m$	EV		

All Studies

m Value	Best Theory
$m < 2.3$	MF
$2.3 \leqslant m < 6.6$	PT
$6.6 \leqslant m < 38.6$	EU
$38.6 \leqslant m$	EV

theories depending on the multiplier m attached to degrees of freedom. For all the studies combined the menu is (in order of increasing taste for parsimony): Mixed fanning, prospect theory, EU, and EV.[26,27]

A theory on one of the Table XIII menus has passed a more rigorous test than simply being undominated. For example, EU is never dominated, but it is never selected as the best model when lottery supports are different. EU is selected for a broad range of m values for lotteries in the interior of the triangle (where lottery supports are the same); the Schwarz criterion selects EU in every case in the interior. Nevertheless, even in the triangle interior EU's performance is not beyond reproach; EU's p value for large gains is small, and in some cases EU is only the best theory when the multiplier values are large. A

[26] The results are changed very little if indifference responses from the BKJ and Harless studies are included (see footnote 17). If each indifference response is assigned to maximize the likelihood function for each individual theory, then the menu of best theories over all studies is: MF for $m < 2.4$, PT for $2.4 \leqslant m < 6.6$, EU for $6.6 \leqslant m < 38.7$, and EV for $38.7 \leqslant m$.

[27] EV may also be interpreted as predicting equal pattern proportions for studies with mean-preserving spread choices (see footnote 6). Aggregated over all studies, this interpretation of EV generates a chi-squared statistic of 1502.7 with 129 degrees of freedom for a p value of $1E-166$. In Table XIII the values of m at which EV becomes best are: 14.7, 5.6, 8.0 (boundary), 80.3, 15.0, 15.7 (interior), and 16.7 instead of 38.6 (all studies).

refinement of prospect theory or a new theory which captured the boundary-interior differences could come within striking distance of EU even in the triangle interior.[28] The m value menu provides no support for betweenness as a middle ground between independence and nonlinear indifference curves: stick by independence (EV on the boundary, EU in the interior) or abandon independence and its weaker cousin betweenness for, say, prospect theory.

4. CONCLUSION

Daniel Bernoulli resolved the St. Petersburg Paradox by replacing mathematical expectation with moral expectation. But Nicholas Bernoulli, who formulated the St. Petersburg paradox, never accepted his cousin's solution, believing that there should be a single fair price for the game. As Stigler (1950) writes, economists may find it surprising that Nicholas Bernoulli and eighteenth century mathematicians believed that the St. Petersburg Paradox could only be "solved" by finding a single price for the game. Might future economists find it peculiar that twentieth century economists held firmly to EU in the face of the Allais paradox and other violations? Stigler's analysis of the development of utility theory through the beginning of this century leads him to three criteria for successful theories: generality, congruence with reality (or fit, in our terms), and manageability. We mention each of these criteria in summarizing the main points of this paper.

There have been many experimental studies comparing EU with competing theories of decision making under risk. Many of these studies use a similar format: Subjects are given several pairwise choices between choices (for example, picking the riskier gamble from one pair implies picking the riskier gamble in another pair). Various theories can then be cast as predictions about patterns of choices that should be observed. We note two important features of our approach: First, our goal is to discriminate among theories which attempt to *describe* actual choices; we have nothing to say about the normative appeal of EU or its generalizations. Second, the generalizability of our results is limited to the extent that naturally-occurring choices are different from lotteries with well-specified probabilities of monetary outcomes.

We conducted analyses of 23 data sets containing nearly 8,000 choices and 2,000 choice patterns, and aggregated the results. We draw several specific conclusions.

(1) All the theories are rejected by a chi-squared test. For every theory there is systematic variation in excluded patterns which could, in principle, be explained by a more refined theory.

[28] As explained in footnote 7, we do not include prospective reference theory (PRT) in the main analysis because none of the studies included here adequately test the predictions of the theory. It is notable that PRT coincides with EU for choices between gambles in the triangle interior, where EU predicts most accurately, but not for boundary gambles for which EU predicts poorly. PRT illustrates how capturing the boundary-interior distinction can improve the predictive utility of a theory. If we were to include PRT in the menus in Table XIII, PRT would not appear on any of the menus for gambles on the boundary of the triangle (PRT is dominated for large gains and for small gains). But since PRT is identical to EU in the interior, over all studies PRT would be selected over PT for $m \geq 4.4$ and EU would be selected over PRT for $m \geq 9.3$.

(2) There is room for improvement in two directions. Some theories, like EU and WEU, are too lean: They could explain the data better by allowing a few more common patterns. Other theories, such as mixed fanning and rank-dependent EU, are too fat: They allow a lot of patterns which are rarely observed. Our analyses provide theorists with a way to diagnose empirical shortcomings of current theories, and perhaps inspiration for new theorizing.

(3) There are dramatic differences between theory accuracy when the gambles in a pair have different support (they lie on the triangle boundary) and when they have the same support (they lie in the triangle interior). EU predicts poorly when support is different, and predicts well when support is the same. The transition from the boundary to the interior implies adding support, typically a small probability of an outcome. Therefore, the accuracy of EU in the interior and its inaccuracy on the boundary suggests that nonlinear weighting of small probabilities is empirically important in explaining choice behavior. This conclusion has been suggested before, but is confirmed dramatically by our analysis. Indeed, Morgenstern (1979) himself accepted that EU had limited applicability when probabilities were low:

> "Now the von Neumann-Morgenstern utility theory, as any theory, is only an approximation to an undoubtedly much richer and far more complicated reality than that which the theory describes in a simple manner.
> ...one should now point out that the domain of our axioms on utility theory is also restricted. Perhaps we should have pointed that out, instead of assuming that this would be understood *ab ovo*. For example, the probabilities used must be within certain plausible ranges and not go to 0.01 or even less to 0.001, then to be compared with other equally tiny numbers such as 0.02, etc. Rather, one imagines that a normal individual would have some intuition of what 50:50 or 25:75 means, etc." (Morgenstern (1979, p. 178).)

(4) The broadest conclusion of our analysis is that there are some losers among competing theories, and some winners. Losers include general theories which rely on betweenness rather than independence, and theories which assume fanning in throughout the triangle; those theories are dominated by other theories which use fewer free parameters and are more accurate. There is some irony here: Some of the theories we test were developed after theorists had seen some of the data sets—these include mixed fan (which we concocted), linear mixed fan or disappointment-aversion theory, lottery dependent utility, etc. It is clear that the development of linear mixed fanning, say, was influenced by data we use to test linear mixed fan. Instead of presenting a problem, the results testify to the power of our approach: We are able to reject some theories using the same data which were taken as inspiration, or support, for developing the theory in the first place.

We cannot declare a single winner among theories—much as we cannot declare a best ice cream or university—because the best theory depends on one's tradeoff between parsimony and fit. But suppose a researcher can specify a single parameter expressing the price of precision, or the reduction in goodness-of-fit (measured by a chi-squared statistic) necessary to justify allowing an extra free parameter. (Some statistical criteria suggest what this price should

be.) We construct a menu of theories which are best at each price-of-precision; researchers can then use the menu to decide which theory to adopt, depending on the price they are willing to pay.

When lotteries have different support, there is *never* a price-of-precision which justifies using EU; anyone who values parsimony enough to use EU over all the generalizations should use EV instead of EU. Combining all the studies (see the bottom of Table XIII), the menu of best theories is: mixed fanning, prospect theory, EU, and EV. Statistical criteria suggest various prices of precision which favor either mixed fanning or EU; the middle ground between high and low prices favors prospect theory.

We cannot give a more definitive answer to the question of which theory is best because people use theories for different purposes. A researcher interested in a broad theory, to explain choices by as many people as possible, cares less for parsimony and more for accuracy; she might choose mixed fanning or prospect theory. A decision analyst who wants to help people make more coherent decisions, by adhering to axioms they respect but sometimes wander from, might stick with EU or EV. A mathematical economist who uses the theory as a brick to build theories of aggregate behavior may value parsimony more highly; she might choose EU or EV (though she should never use EU when choices involve gambles with different support).

However, an historical parallel described by Stigler (1950) may be instructive for those who cling to EU:

> "Economists long delayed in accepting the generalized utility function because of the complications in its mathematical analysis, although no one (except Marshall) questioned its realism... Manageability should mean the ability to bring the theory to bear on specific economic problems, not ease of manipulation. The economist has no right to expect of the universe he explores that its laws are discoverable by the indolent and the unlearned. The faithful adherence for so long to the additive utility function strikes one as showing at least a lack of enterprise" (Stigler (1950, pp. 393–394).)

The pairwise-choice studies suggest that violations of EU are robust enough that modeling of aggregate economic behavior based on alternatives to EU is well worth exploring. So far there have been relatively few such efforts.[29] Ultimately, most of the payoff for economics will come from replacing EU in models of individual behavior with more accurate descriptive principles or a single formal theory. Our results suggest which replacements are most promising, and which modifications of the currently available theories are most productive.

We see our paper as summarizing a chapter in the history of empirical studies of risky choice. We think the weight of evidence from recent studies with multiple pairwise choices, when aggregated across those studies, is sufficiently great that new pairwise-choice studies are unlikely to budge many basic conclusions—the statistical value-added of more such studies is low (compared to the

[29] Epstein (1990) reviews some recent efforts by economists.

value-added of new approaches). However, this sweeping conclusion leans heavily on the assumption that different studies are completely independent (which they likely are not). If studies are highly dependent then our results are overstated, and there may still be substantial value in using the pairwise-choice paradigm to exploring new domains of gambles (e.g., gambles over losses, gambles with many possible outcomes, gambles with very low probabilities); more data *could* change the way at least some theories are ranked.

If our analysis closes the chapter on pairwise-choice empirics (or summarizes the much that we know so far), then it opens new chapters as well—particularly, a chapter devoted to combining structural explanations of choice problems with more sophisticated theories of errors. Empirical studies fitting individual non-EU functions and parameters to subjects are useful and relatively rare (see Daniels and Keller (in press), Hey and Di Cagno (1990), Tversky and Kahneman (1992)). Studies that test axioms directly—e.g., Wakker, Erev, and Weber (1993) test comonotonic independence, the crucial ingredient in rank-dependent approaches—are useful too. Function-fitting, and our approach, both allow heterogeneous preferences. (The fact that estimated pattern proportions are fairly even across patterns suggest there *is* substantial heterogeneity.) For analytical tractability, it is often useful to assume homogeneity (in representative-agent models); then the sensible empirical question is which single theory, and which precise parameter values, fits everyone's choices best (see Camerer and Ho (in press)).

Finally, our general method could be applied in other domains. For example, various noncooperative solution concepts permit different sets of choices in games. Theories could be characterized as restrictions on allowable patterns of choices, and the distribution of patterns could be explicitly connected through an error rate. For example, McKelvey and Palfrey (1992) apply a similar error theory to fit various equilibrium concepts to experimental data on the "centipede" game, and to test restrictions imposed by different concepts; El-Gamal and Grether (1993) apply a similar analysis to experimental data on probability judgments. Most importantly, our method would allow one to judge which concepts, like Nash equilibrium or its various refinements (and coarsenings), best trade off parsimony and accuracy. A similar method could be applied to compare solution concepts in cooperative games. The discussion above shows how our method uses more information and hence is more powerful than methods which judge theories only by the percentage of consistent responses (e.g., Selten (1987)).

Dept. of Economics, Virginia Commonwealth University, Richmond, VA 23284, U.S.A.

and

Graduate School of Business, University of Chicago, 1101 E. 58th St., Chicago, IL 60637, U.S.A.

Manuscript received February, 1992; final revision received February, 1994.

REFERENCES

AITKIN, M. (1991): "Posterior Bayes Factors," *Journal of the Royal Statistical Society, Series B*, 53, 111–142.

AKAIKE, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proceedings of the 2nd International Symposium on Information Theory*, ed. by N. Petrov and F. Csadki. Budapest: Akademiai Kiado.

ALLAIS, M. (1953): "'Le Comportement de l'Homme Rationnel devant le Risque, Critique des Postulates et Axiomes de l'Ecole Americaine," *Econometrica*, 21, 503–546.

BATTALIO, R. C., J. H. KAGEL, AND K. JIRANYAKUL (1990): "Testing Between Alternative Models of Choice Under Uncertainty: Some Initial Results," *Journal of Risk and Uncertainty*, 3, 25–50.

BECKER, J. L., AND R. SARIN (1987): "Lottery Dependent Utility," *Management Science*, 33, 1367–1382.

CAMERER, C. F. (1989): "An Experimental Test of Several Generalized Utility Theories," *Journal of Risk and Uncertainty*, 2, 61–104.

——— (1992): "Recent Tests of Generalizations of EU Theories," in *Utility: Theories, Measurement, and Applications*, ed. by W. Edwards. Dordrecht: Kluwer.

CAMERER, C. F., AND T.-H. HO (in press): "Violations of the Betweenness Axiom and Nonlinearity in Probability," *Journal of Risk and Uncertainty*, forthcoming.

CHEW, S. H. (1983): "A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox," *Econometrica*, 51, 1065–1092.

——— (1989): "Axiomatic Utility Theories with the Betweenness Property," *Annals of Operations Research*, 19, 273–298.

CHEW, S. H., AND L. G. EPSTEIN (1990): "A Unifying Approach to Axiomatic Non-Expected Utility Theories," *Journal of Economic Theory*, 49, 207–240.

CHEW, S. H., L. G. EPSTEIN, AND U. SEGAL (1991): "Mixture Symmetry and Quadratic Utility," *Econometrica*, 59, 139–163.

CHEW, S. H., E. KARNI, AND Z. SAFRA (1987): "Risk Aversion in the Theory of Expected Utility with Rank Dependent Probabilities," *Journal of Economic Theory*, 42, 370–381.

CHEW, S. H., AND K. R. MACCRIMMON (1979): "The HILO Structure and the Allais Paradox," University of British Columbia Working Paper.

CHEW, S. H., AND W. WALLER (1986): "Empirical Tests of Weighted Utility Theory," *Journal of Mathematical Psychology*, 30, 55–62.

CONLISK, J. (1989): "Three Variants on the Allais Example," *American Economic Review*, 79, 392–407.

CRAWFORD, V. P. (1988): "Stochastic Choice with Quasiconcave Preference Functions," Department of Economics 88-28, University of California, San Diego.

DANIELS, R. L., AND L. R. KELLER (in press): "An Experimental Evaluation of the Descriptive Validity of Gamble Dependent Utility," *Journal of Risk and Uncertainty*, forthcoming.

DEKEL, E. (1986): "An Axiomatic Characterization of Preferences Under Uncertainty: Weakening the Independence Axiom," *Journal of Economic Theory*, 40, 304–318.

EL-GAMAL, M., AND D. GRETHER (1993): "Uncovering Behavioral Strategies: Likelihood-based Experimental Data-mining," Caltech Division of Social Sciences and Humanities Working Paper.

EPSTEIN, L. G. (1990): "Behavior Under Risk: Recent Developments in Theory and Applications," University of Toronto Department of Economics Working Paper.

FISHBURN, P. C. (1982): "Nontransitive Measurable Utility," *Journal of Mathematical Psychology*, 26, 31–67.

——— (1983): "Transitive Measurable Utility," *Journal of Economic Theory*, 31, 293–317.

——— (1988): *Nonlinear Preference and Utility Theory*. Baltimore, MD: Johns Hopkins Press.

GREEN, J. R., AND B. JULLIEN (1988): "Ordinal Independence in Nonlinear Utility Theory," *Journal of Risk and Uncertainty*, 1, 355–387. (Erratum, vol. 2, p. 119.)

GUL, F. (1991): "A Theory of Disappointment Aversion," *Econometrica*, 59, 667–686.

HARLESS, D. W. (1992): "Predictions About Indifference Curves Inside the Unit Triangle: A Test of Variants of Expected Utility Theory," *Journal of Economic Behavior and Organization*, 18, 391–414.

——— (1993): "Experimental Tests of Prospective Reference Theory," *Economic Letters*, 43, 71–76.

HEY, J. D., AND D. DICAGNO (1990): "Circles and Triangles: An Experimental Estimation of Indifference Lines in the Marschak-Machina Triangle," *Journal of Behavioral Decision Making*, 3, 279–306.

HEY, J., AND C. ORME (1993): "Investigating Parsimonious Generalizations of Expected Utility Theory using Experimental Data," University of York Centre for Experimental Economics Working Paper.

KAHNEMAN, D., AND A. TVERSKY (1979): "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, 47, 263–291.

KLEIN, R. W., AND S. J. BROWN (1984): "Model Selection When There is 'Minimal' Prior Information," *Econometrica*, 52, 1291–1312.

LICHTENSTEIN, S., AND P. SLOVIC (1971): "Reversal of Preferences Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, 89, 46–55.

MACHINA, M. J. (1982): "'Expected Utility' Analysis Without the Independence Axiom," *Econometrica*, 50, 277–323.

——— (1985): "Stochastic Choice Functions Generated from Deterministic Preferences over Lotteries," *Economic Journal*, 95, 575–594.

——— (1987): "Choice Under Uncertainty: Problems Solved and Unsolved," *Journal of Economic Perspectives*, 1, 121–154.

MARSCHAK, J. (1950): "Rational Behavior, Uncertain Prospects, and Measurable Utility," *Econometrica*, 18, 111–141.

MCKELVEY, R., AND T. PALFREY (1992): "An Experimental Investigation of the Centipede Game," *Econometrica*, 60, 803–836.

MORGENSTERN, O. (1979): "Some Reflections on Utility Theory," in *EU Hypotheses and the Allais Paradox*, ed. by M. Allais and O. Hagen. Dordrecht: D. Reidel.

NEILSON, W. S. (1992a): "A Mixed Fan Hypothesis and its Implications for Behavior Towards Risk," *Journal of Economic Behavior and Organization*, 19, 197–212.

——— (1992b): "Some Mixed Results on Boundary Effects," *Economics Letters*, 39, 275–278.

NELDER, J. A., AND R. W. M. WEDDERBURN (1972): "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

PRELEC, D. (1990): "A 'Pseudo-Endowment' Effect and its Implications for Some Recent Non-EU Models," *Journal of Risk and Uncertainty*, 3, 247–259.

QUIGGIN, J. (1982): "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization*, 3, 323–343.

——— (1993): "Testing Between Alternative Models of Choice Under Uncertainty: Comment," *Journal of Risk and Uncertainty*, 6, 161–164.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

SEGAL, U. (1987): "Some Remarks on Quiggin's Theory of Anticipated Utility," *Journal of Economic Behavior and Organization*, 8, 145–154.

——— (1989): "Anticipated Utility: A Measure Representation Approach," *Annals of Operations Research*, 19, 359–373.

SELTEN, R. (1987): "Equity and Coalition Bargaining in Experimental Three Person Games," in *Laboratory Experiments in Economics: Six Points of View*, ed. by A. Roth. Cambridge: Cambridge University Press.

——— (1991): "Properties of a Measure of Predictive Success," *Mathematical Social Sciences*, 21, 153–200.

SMITH, A. F. M., AND D. J. SPIEGELHALTER (1980): "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society, Series B*, 42, 213–220.

SMITH, V., AND J. WALKER (1993): "Monetary Rewards and Decision Cost in Experimental Economics," *Economic Inquiry*, 31, 245–261.

SOPHER, B., AND G. GIGLIOTTI (1990): "A Test of Generalized Expected Utility," Department of Economics, Rutgers University.

STARMER, C., AND R. SUGDEN (1989): "Probability and Juxtaposition Effects: An Experimental Investigation of the Common Ratio Effect," *Journal of Risk and Uncertainty*, 2, 159–178.

STIGLER, G. J. (1950): "The Development of Utility Theory. II," *Journal of Political Economy*, 58, 373–396.

TVERSKY, A., AND D. KAHNEMAN (1992): "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty*, 5, 297–323.

VISCUSI, K. W. (1989): "Prospective Reference Theory: Toward An Explanation of the Paradoxes," *Journal of Risk and Uncertainty*, 2, 235–264.

WAKKER, P., I. EREV, AND E. WEBER (1993): "Comonotonic Independence: The Critical Test Between Classical and Rank-Dependent Utility Theories," University of Chicago Center for Decision Research Working Paper.

YAARI, M. E. (1987): "The Dual Theory of Choice Under Risk," *Econometrica*, 55, 95–115.